Stichprobenqualität in Bevölkerungsumfragen

Faulbaum, Frank (Ed.); Wolf, Christof (Ed.)

Veröffentlichungsversion / Published Version
Konferenzband / conference proceedings

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Stichprobenqualität in Bevölkerungsumfragen

Frank Faulbaum, Christof Wolf (Hrsg.)

Tagungsberichte, Band 12

InformationsZentrum Sozialwissenschaften

GESIS

Stichprobenqualität in Bevölkerungsumfragen

Tagungsberichte

Herausgegeben vom Informationszentrum Sozialwissenschaften (IZ)
der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI), Bonn.
Band 12

Das IZ ist Mitglied der Gesellschaft Sozialwissenschaftlicher
Infrastruktureinrichtungen e.V. (GESIS).
Die GESIS ist Mitglied der Leibniz-Gemeinschaft.

Stichprobenqualität in Bevölkerungsumfragen

Frank Faulbaum, Christof Wolf (Hrsg.)

Tagungsberichte, Band 12

Informationszentrum Sozialwissenschaften, Bonn 2006

Bibliographische Information Die Deutsche Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliothek; detaillierte bibliographische Daten sind im Internet über www.ddb.de abrufbar.

ISBN-10 3-8206-0156-2
ISBN-13 978-3-8206-0156-5

Herausgeber, Druck und Vertrieb:
Informationszentrum Sozialwissenschaften
Lennéstraße 30, 53113 Bonn
Tel.: 02 28 - 22 81 - 0
Printed in Germany

Inhalt

Einleitung

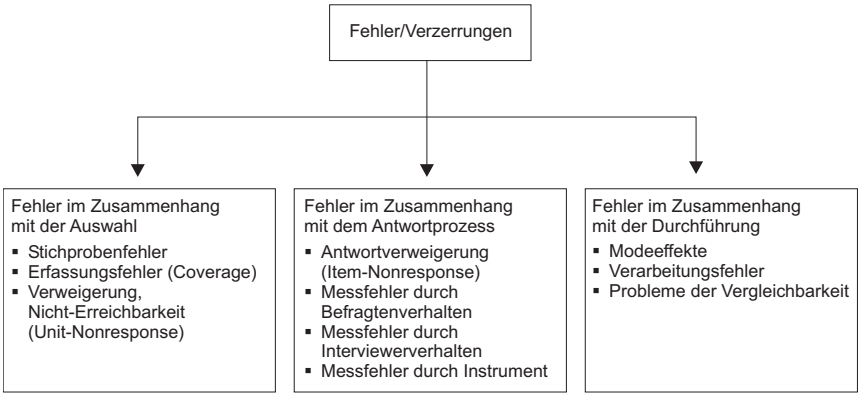
Frank Faulbaum, Christof Wolf

Der vorliegende Band enthält Beiträge einer Tagung über „Stichprobenqualität in Bevölkerungsumfragen“, die gemeinsam von der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (ASI) und der Sektion „Methoden der Empirischen Sozialforschung“ der Deutschen Gesellschaft für Soziologie am 14./15. Oktober 2005 in Berlin veranstaltet wurde.

Bevölkerungsumfragen sind nicht nur unverzichtbares Instrument sozialwissenschaftlicher Forschung, sondern gehören inzwischen zu den unumstrittenen Hilfsmitteln und wohl etablierten Instrumenten wirtschafts-, bildungs-, kultur- und sozialpolitischer Entscheidungsvorbereitung. Politische Entscheidungen in hoch entwickelten Gesellschaften erfordern, nicht nur zur rechtzeitigen Prognose krisenhafter Entwicklungen, sondern auch zur Erarbeitung kurz- und mittelfristiger Planungsunterlagen die systematische Sammlung von Erkenntnissen über Veränderungen in Wirtschaft und Gesellschaft. Längerfristige Planungen bedürfen dabei immer wieder der zwischenzeitlichen empirischen Überprüfung. Zum Teil werden diese Erhebungen, wie etwa im Fall des Mikrozensus, auf gesetzlicher Grundlage vom Staat, vertreten durch das Statistische Bundesamt und die statistischen Ämter, selbst durchgeführt, zum Teil als Forschungsaufträge an akademische oder privatwirtschaftliche Institutionen vergeben. Auch auf regionaler und kommunaler Ebene, auf der Ebene von Städten und Gemeinden, haben sich Umfragen inzwischen mehr und mehr zu einem methodischen Standardinstrument der Sozialberichterstattung, der wirtschaftlichen und sozialen Dauerbeobachtung, der Vorbereitung und Evaluation kultur-, sozial- und arbeitsmarktpolitischer kommunaler Maßnahmen und der Untersuchung der Akzeptanz kommunaler Entscheidungen entwickelt.

Die Bedeutung von Bevölkerungsumfragen als wichtige politische Entscheidungsgrundlage und als wichtiger Lieferant von Daten für die wissenschaftliche Forschung stellt besondere Anforderungen an die Umfrage*qualität*. Mit Recht weist die Denkschrift der Deutschen Forschungsgemeinschaft (vgl. Kaase 1999: 96) darauf hin, dass eine Methodologie der Qualitätsbewertung von Umfragen eine ganzheitliche Perspektive einnehmen sollte. In diesem Sinne wird die Prozessqualität von Umfragen immer wieder thematisiert (vgl. z.B. Lyberg et al. 1997; Biemer & Lyberg 2003). Der Begriff der Umfragequalität lässt sich weiter präzisieren, wenn man sich die unmittelbaren Ziele einer Umfrage vergegenwärtigt. Ziel einer Umfrage ist es in der Regel Aussagen über Parameter (Erwartungs-

werte, Varianzen, etc.) in einer genau definierten Population auf der Basis von Stichprobendaten zu machen. In einer Umfrage werden Messungen zur Schätzung von Quantitäten in einer Population vorgenommen, z.B. des Prozentsatzes der Wähler einer bestimmten Partei, des Prozentsatzes der Käufer eines bestimmten Produkts, des Prozentsatzes der Familien, die in einer bestimmten Großstadt wohnen, des mittleren Einkommens, etc. Eine Möglichkeit, die Qualität solcher *Umfragemessungen* (survey measurements) zu definieren, wäre, die tatsächlichen Messungen mit denen zu vergleichen, die man erhalten würde, wenn die Umfrage unter idealen Bedingungen hätte durchgeführt werden können, d.h., wenn die Ziele der Messung optimal hätten umgesetzt werden können (vgl. Ditto 1997). Diese idealen Bedingungen können durch verschiedene Arten von Fehlern verzerrt werden. Abbildung 1 gibt einen Überblick über die Fehler, mit denen in Umfragen zu rechnen ist.



Quelle: Nach Weiberg (2005: 19)

Abbildung 1: Fehlerarten in Umfragen

Die in diesem Band versammelten Beiträge konzentrieren sich thematisch vor allem auf die Qualität der Stichprobe und damit vor allem auf die Stichprobenfehler und Nichtbeobachtungsfehler (nonobservation errors).

Diese Fehler sind in ihren Entstehungsbedingungen nicht unabhängig vom Typus der Umfrage zu sehen. Umfragen und Methoden der Datenerhebung im Allgemeinen lassen sich grob typisieren nach der Erhebungsart (collection mode; z.B. face-to-face, telefonisch, schriftlich), der Erhebungstechnologie (collection technology; z.B. computerunterstützt, paper&pencil) und nach der Administrationsform (intervieweradministriert, selbstadministriert). Die Zugehörigkeit zu einem dieser Typen hat bestimmte Konsequenzen für die Definierbarkeit von Populationen, Form, Gestalt und Vollständigkeit der Auswahlgrundlage (sampling

frame) sowie auf die verschiedenen Formen des Nichtbeobachtungsfehlers und damit auf die Stichprobenqualität, die wiederum eine wichtige Determinante der Umfragequalität darstellt.

Im Mittelpunkt der Beiträge stehen unterschiedliche Varianten der Stichprobenziehung, deren Entwicklung durch praktische Erwägungen oder durch Veränderungen des Befragtenverhaltens angestoßen wurde, die Entwicklung und Optimierung von Schätzverfahren für verschiedene Erhebungsarten und Erhebungstechnologien sowie die Determinanten systematischer Ausfälle und Möglichkeiten ihrer Reduktion.

So beschäftigen sich die Beitrage von Häder und Gabler, Hoffmeyer-Zlotnik, Blohm, Krügener, Jann und Holst mit den Ziehungsverfahren. Sabine Häder und Siegfried Gabler geben einen Überblick über neue Entwicklungen bei der Ziehung von Telefonstichproben in Deutschland, die durch die zunehmende Handy-Nutzung notwendig werden. In dem Beitrag von Jürgen Hoffmeyer-Zlotnik sowie dem Beitrag von Michael Blohm werden die Qualitätsunterschiede verschiedener Ziehungsverfahren in face-to-face Befragungen untersucht. Sonja Krügener gibt einen Überblick über den aktuellen Stand des Registerzensus als Alternative zur traditionellen Volkszählung. Ben Jann stellt einen Vorschlag für eine effiziente Klumpenstichprobe am Beispiel der Schweiz vor und Christian Holst erläutert die Stichprobenanlage des Ipsos SOWI-Bus.

Die Beiträge von Münnich und Magg sowie von Schnell und Trappmann beschäftigen sich mit Schätzverfahren für bestimmte Formen der Datenerhebung und Umfragedesigns. Ralf Münnich und Kersten Magg befassen sich mit der Schätzqualität im registergestützten Zensus, Rainer Schnell und Mark Trappmann untersuchen die Konsequenzen der Mortalität für die Schätzung der Lebenserwartung im SOEP.

Mit den Determinanten systematischer Ausfälle sowie den Möglichkeiten, systematische Ausfälle zur reduzieren, befassen sich der Beitrag von Baur und der Beitrag von Engel. Nina Baur beschäftigt sich mit den Ausfallgründen bei zufallsgenerierten Telefonstichproben. Uwe Engel gibt einen Überblick über Anreizeffekte in Studien der Markt- und Sozialforschung, also über gewisse Strategien, Ausfalleffekte zu reduzieren.

Der Beitrag von Bernhard Schimpl-Neimanns bezieht sich auf den Beobachtungsfehler bei der Durchführung einer Umfrage. Er befasst sich mit der Datenqualität der Bildungsangaben im Mikrozensus.

Literatur

Biemer, P.P. & Lyberg, L.E. (2003). *Introduction to Survey Quality*. New York: Wiley.

Dippo, C.S. (1997). Survey measurement and process improvement: Concepts and integration. In: Lyberg, L. et al. (eds.). *Survey measurement and process quality*. New York: Wiley, 457-474.

Kaase, M. (Hrsg.) (1999). *Deutsche Forschungsgemeinschaft. Qualitätskriterien der Umfrageforschung:Memorandum*. Berlin: Akademie Verlag

Lyberg, L. et al. (Eds.). (1997). *Survey measurement and process quality*. New York: Wiley.

Weisberg, H.F. (2005). *The total survey error approach*. Chicago: The University of Chicago Press.

Neue Entwicklungen bei der Ziehung von Telefonstichproben in Deutschland

Sabine Häder, Siegfried Gabler

Über 40 Prozent aller Interviews in der Marktforschung werden in Deutschland gegenwärtig telefonisch durchgeführt (vgl. Tabelle 1). Voraussetzung für eine hohe Qualität der Telefonumfragen ist die adäquate Generierung der Stichproben. Als Auswahlrahmen hat sich seit Ende der 1990er Jahre in Deutschland ein bei ZUMA entwickelter Frame (Gabler-Häder-Design) durchgesetzt, der sowohl in das Telefonbuch eingetragene wie auch nicht eingetragene Anschlüsse enthält, die über ein Ortsnetz erreichbar sind (Gabler/Häder 2002). Dieser Auswahlrahmen wurde vom ADM durch für die Schichtung geeignete Merkmale angereichert (ADM-Design).

Tabelle 1: Quantitative Interviews der Mitgliedsinstitute des ADM nach Befragungsart (in %)

	1990	1991	1995	1996	1997	2002	2003	2004
Persönliche Interviews	65	60	60	45	44	33	28	31
Telefon-Interviews	**22**	**30**	**30**	**44**	**40**	**41**	**43**	**44**
Schriftliche Interviews	13	10	10	11	16	21	19	9
Online-Interviews	-	-	-	-	-	5	10	16

Quelle: ADM e.V. (2005)

In den letzten Jahren hat sich allerdings eine Tendenz angedeutet, die die alleinige Nutzung dieses Auswahlrahmens als unzureichend zur Abdeckung der Gesamtheit der Privathaushalte erscheinen lassen könnte: Ein wachsender Anteil der Haushalte ist lediglich über Mobiltelefon erreichbar[1]. Diese Haushalte haben bei telefonischen Umfragen mit Stichproben nach dem Gabler-Häder-Design oder dem ADM-Design keine positive Auswahlchance, sofern sie nicht über eine vir-

1 „Mit 74 Millionen hat die Zahl der Handy-Nutzer zum Halbjahr (2005, d.A.) in Deutschland ein neues Rekordhoch erreicht. Allerdings geht das Wachstum im Mobilfunk teilweise zu Lasten der Sprachtelefonie im Festnetz, weil Handys immer häufiger auch zu Hause eingesetzt werden." (Müller 2005)

tuelle Festnetznummer verfügen (z. B. O2). Damit kann es zu systematischen Verzerrungen in den Stichproben kommen, wenn sich Festnetzhaushalte und Mobilfunkhaushalte hinsichtlich für die Sozialforschung relevanter Merkmale unterscheiden. Deshalb sind Überlegungen über die Integration von Mobilfunkanschlüssen in Telefonstichproben notwendig.

1 Ausbreitung der Mobilfunkhaushalte

Zunächst ist es interessant, sich einen Überblick darüber zu verschaffen, in welchem Maße sich die Haushalte, die nur über Mobilfunk erreichbar sind, in der letzten Zeit ausgebreitet haben. Dazu sind Daten aus Face-to-Face-Befragungen hilfreich, in denen u.a. nach der Erreichbarkeit der Personen über Telefon gefragt wird. So ergeben die Media-Analysen ma 2004 Pressemedien II (n=32.423) bzw. ma 2005 Pressemedien I (n=38.904) hochgerechnet folgende Werte für Personen, die lediglich über Mobiltelefon(e) verfügen (siehe Tabelle 2):

Tabelle 2: Anzahl Personen mit ausschließlichem Mobilfunkanschluss

Media-Analyse	BRD gesamt		Westdeutschland		Ostdeutschland	
	n	in %	*n*	in %	*n*	in %
2004	1.913	3,0	1.238	2,4	675	5,1
2005	2.170	3,4	1.431	2,8	740	5,6

Demnach sind gegenwärtig 3,4% der Personen nur über ein Handy erreichbar. Dies wäre zunächst noch kein Anlass, das unmittelbar bevorstehende Ende der Festnetzstichproben befürchten zu müssen. Schwierig wird die Situation aber durch die sozialstrukturellen Unterschiede in der Verbreitung von Mobiltelefonen. So ist der Anteil von „Nur-Mobilfunk-Erreichbaren“ in Ostdeutschland höher als in Westdeutschland, bei Frauen niedriger als bei Männern, in Einpersonenhaushalten höher als in Mehrpersonenhaushalten sowie bei Arbeitslosen und in Ausbildung Befindlichen höher als bei Berufstätigen und Rentnern. Andere Untersuchungen kommen zu ähnlichen Ergebnissen. TNS Infratest ermittelte in seinem Jahres-BUS 2004 Anteile von 5,9% der Personen bzw. 7,4% der Haushalte, die ausschließlich über ein Handy verfügen. Die gefundenen sozialstrukturellen Unterschiede bestätigen die Ergebnisse aus der Media-Analyse.

Allerdings ist es denkbar, dass der in Face-to-Face-Befragungen ermittelte Anteil von Haushalten, die mit dem Auswahlrahmen für Festnetzstichproben nicht erreicht werden können, überschätzt wird. Auf die Frage, ob sie zu Hause nur über Mobiltelefon erreichbar seien, antworten Homezone-Kunden korrekterwei-

se mit „ja“.[2] Homezone erlaubt es, mit einem Handy in einer festgelegten Zone zum Festnetzpreis zu telefonieren. Wenn diese Zone verlassen wird, schaltet das Handy auf Mobilfunktarif. Wichtig im Zusammenhang mit der Stichprobenziehung für Telefonumfragen ist nun, dass bei diesem Angebot in der Regel eine Festnetznummer vergeben wird, d.h. diese im bisher benutzten Auswahlrahmen enthalten ist. Insofern können diese Teilnehmer auch gegenwärtig schon erreicht werden. Dadurch dürfte der Anteil nicht erreichbarer Personen bzw. Haushalte gegenwärtig geringer sein als oben angegeben. Eventuell sollte deshalb die bisher übliche Abfrage dadurch ersetzt werden, dass erkundet wird, ob die befragte Person über ein Telefon mit Ortsnetzkennzahl erreichbar ist.

Dennoch: Durch die Einführung von UMTS und anderen drahtlosen Datenübertragungsmöglichkeiten wird zukünftig der Zwang zum Festnetzanschluss als Voraussetzung für den Internet-Anschluss entfallen. Damit ist ein weiterer Anstieg der Mobilfunkhaushalte zu erwarten. Die Stichprobenstatistiker in Deutschland sind deshalb aufgerufen, eine Lösung für die Integration von Mobilfunknummern in Telefonstichproben zu finden.

2 Konstruktion von Auswahlrahmen für Mobilfunkrufnummern

Mobilfunkrufnummern sind grundsätzlich anders organisiert als Festnetzrufnummern. Die Vorwahlbereiche sind jeweils Netzbetreibern – nicht Regionen - zugeordnet. Rufnummerngruppen im Bereich 015x bis 017x sind einzelnen Mobilfunkanbietern zugeteilt. Bei einem Wechsel zu einem anderen Anbieter können durch die Rufnummermitnahme die Nummern dann aber bei einem anderen Provider liegen.

Das Problem besteht nun in der Konstruktion eines geeigneten Auswahlrahmens für Mobilfunknummern. Nahe liegend wäre zunächst, ihn analog zum Festnetzstichprobenauswahlrahmen gestützt auf die Telefonbucheinträge zu entwickeln. Da nur ein äußerst geringer Anteil der Mobilfunknummern im Telefonbuch enthalten ist, ist allerdings zu befürchten, dass mit dem herkömmlichen Vorgehen ein zu geringer Teil der Blöcke im Auswahlrahmen enthalten wäre[3]. Ein solcher Frame wurde bei ZUMA konstruiert und einem Gütetest unterzogen (siehe Tabelle 3).

2 Dies ist die übliche Abfrage in Interviews, um den Anteil von Nur-Mobilfunkhaushalten schätzen zu können.

3 Im hier vorgestellten Auswahlrahmen waren 015x-Nummern noch nicht enthalten, wohl aber in den Teststichproben.

Tabelle 3: Überdeckung von Mobilfunknummern durch den ZUMA-Auswahlrahmen

		BLÖCKE					
	Umfang	100er		1000er		10000er	
	n	abs	in %	abs	in %	abs	in %
ZUMA-Stichprobe	140	79	56,4	107	76,4	131	93,6
Studenten-Stichprobe	558	294	52,7	427	76,5	506	90,7

Die Überdeckung des Auswahlrahmens mit dem vorhandenen Mobilfunknummernraum bei der Verwendung von 100er-Blöcken ist mangelhaft. Nur etwas mehr als die Hälfte der Mobilfunknummern in den Stichproben war auch im Auswahlrahmen enthalten. Selbst bei der Verwendung von 1000er-Blöcken ergab sich kein befriedigendes Ergebnis: Lediglich 76% der Nummern aus den Teststichproben waren im Auswahlrahmen zu finden. Deshalb erscheint es bei der Konstruktion des Auswahlrahmens sinnvoll zu sein, alle möglichen Ziffernfolgen innerhalb der Vorwahlbereiche zu generieren. Dieser Gedanke wird von BIK Aschpurwis+Behrens verfolgt. Dabei wird davon ausgegangen, dass die Effizienz (Hitrate) von Stichproben, die auf einem durch Random Digit Dialing erzeugten Auswahlrahmen basieren, erhöht werden kann. Dies wäre prinzipiell möglich, wenn Kenntnis über nicht genutzte Rufnummerngassen vorläge. Auskunft darüber ist allerdings nicht über die Bundesnetzagentur zu bekommen, da die Vergabe der Rufnummern vollständig in der Verantwortung der Netzbetreiber erfolgt. Hier sind weitere Recherchen notwendig.

3 Probleme beim Einsatz eines Auswahlrahmens für Mobilfunkrufnummern

Die Konstruktion eines geeigneten Auswahlrahmens stellt also an sich ein lösbares Problem dar. Da jedoch nicht davon auszugehen ist, dass in naher Zukunft eine vollständige Abdeckung der Haushalte mit Handys erreicht wird, die den Verzicht auf Festnetznummern erlauben würde, müssen die Auswahlrahmen für Mobilfunk- und Festnetzrufnummern parallel verwendet werden. Bei der Kombination beider Auswahlrahmen sind aber einige Schwierigkeiten zu erwarten. Diese sollen im Folgenden kurz angerissen werden:

a) Gewichtung

Festnetzstichproben sind als Haushaltsstichproben aufzufassen, d.h. bei Bevölkerungsbefragungen muss eine Transformation auf Personenebene erfolgen, die die Zahl der zur Grundgesamtheit gehörenden Personen im Haushalt berücksichtigt.

Bei Mobilfunkstichproben handelt es sich jedoch eher um Personenstichproben. Hier könnte die Transformationsgewichtung unterbleiben. Ein Modell für die Kombination beider Stichproben ist zu entwickeln.

Weiterhin ist problematisch, dass im Interview die Zahl der Festnetz- und Mobilfunkanschlüsse zu erfassen ist, um die sich daraus ergebenden unterschiedlichen Inklusionswahrscheinlichkeiten ausgleichen zu können. Allerdings zeigen Erfahrungen, dass die diesbezüglichen Angaben der Befragten nicht immer valide sind. Dies kann zu Verzerrungen führen. Darüber hinaus erfordert die exakte Abfrage der Zahl der verschiedenen Telefonanschlüsse im Interview relativ viel Zeit und ihre Relevanz ist für den Befragten nicht unmittelbar einsehbar. Dies kann zu Problemen bei der Ausschöpfung der Stichprobe führen.

b) Regionale Zuordnung

Das ADM-Design sieht eine regionale Schichtung vor. Diese ist z.B. für die Media-Analyse essentiell, da hier verschiedene Regionen überproportional erhoben werden müssen. Für Festnetzstichproben wurde ein Modell entwickelt, das dieses Problem löst. Bei Mobilfunkanschlüssen ist dagegen eine regionale Verortung der Anschlüsse aus den Rufnummern nicht möglich. Es ist unklar, wie hier vorzugehen ist.

c) Ausschöpfungsberechnung

Bei Mobilfunkrufnummern werden, wenn kein Anschluss zustande kommt, bestimmte Digitalcodes zurückgesendet. Diese sind aber von Provider zu Provider unterschiedlich und nicht immer eindeutig identifizierbar. Das erschwert die Ausschöpfungsberechnung, da jeweils nicht zweifelsfrei entschieden werden kann, ob es sich um einen geschalteten Anschluss handelt, dessen Inhaber nur nicht erreichbar ist, oder ob der Anschluss nicht vergeben ist. An der TU Dresden soll deshalb getestet werden, ob aufgrund der Reaktion auf eine Ankündigungs-SMS entschieden werden kann, ob der Anschluss geschaltet ist oder nicht.

d) Private / geschäftliche Mobilfunknutzung

In den Telefonstichproben nach dem Gabler-Häder- und dem ADM-Design ist die Repräsentanz der Bevölkerung in Privathaushalten das erklärte Ziel. Geschäftlich genutzte Festnetznummern gehören somit nicht zur adäquaten Auswahlgrundlage und werden deshalb bei der Stichprobenauswahl und/oder im Telefonkontakt ausgesteuert. Eine Auswahlgrundlage für Mobilfunknutzer beinhaltet das gleiche Problem in weitaus größerem Maß. Hier fehlen erstens die zur Sonderbehandlung geschäftlicher Anschlüsse geeigneten Informationen aus den Einträgen wie z.B. so genannte „Bindestrichnummern“, die auf einen Firmenanschluss schließen lassen. Zweitens gibt es „Geschäftshandys“, die an Mitarbeiter

zeitweilig oder dauerhaft vergeben werden. Fraglich ist, ob derartige Handys für Befragungen genutzt werden können.

e) Kontaktsituation

Während die Nutzung des Festnetzanschlusses – lässt man Weiterleitungsmöglichkeiten außer Acht, die heute noch wenig verbreitet sind – auf die häuslichen Räume beschränkt ist, ist ein wesentliches Merkmal der Mobilfunknutzung die ständige auch aushäusige Erreichbarkeit; in welchen Situationen heute überall das Handy genutzt wird, kann man jeden Tag in seinem Umfeld beobachten. Bei vielen dieser Situationen kann man sich nur schwer vorstellen, dass die Kontaktaufnahme für ein Interview erfolgreich wäre, oder – bei zunächst positivem Kontakt – das Interview zu verlässlichen Daten führen würde.

f) Mode effects

Selbst wenn die vorgenannten Probleme gelöst sein werden, besteht noch Unsicherheit darüber, welche Mode effects bei der Befragung via Handy auftreten können.

4 Ausblick

Am 1. März 2005 fand bei ZUMA das erste Treffen der Arbeitsgruppe MOBILSAMPLE statt. Diese Arbeitsgruppe, bestehend aus Wissenschaftlern in der akademischen und kommerziellen Sozialforschung, wird an der Lösung der oben genannten Probleme arbeiten. Die Ergebnisse ihrer Forschung werden sie auf den Treffen der Arbeitsgruppe vorstellen. Das zweite Treffen fand am 21. Februar 2006 statt. Um die Forschungsergebnisse auch der Profession bekannt zu machen, sind Veröffentlichungen in geeigneten Zeitschriften (wie z.B. ZUMA Nachrichten) und Präsentationen auf der Tagung am 21. November 2006 bei ZUMA vorgesehen (vgl. Mitteilung über die Gründung der Arbeitsgruppe MOBILSAMPLE, ZUMA-Nachrichten 56: 111-116).

In Kooperation des Lehrstuhls für Methoden der Empirischen Sozialforschung der TU Dresden und ZUMA wurde ein Antrag auf Sachbeihilfe bei der Deutschen Forschungsgemeinschaft gestellt, der inzwischen auch genehmigt wurde. Geplant ist, aus diesen Mitteln eine Bevölkerungsbefragung zu finanzieren, bei der sowohl Festnetz- als auch Mobilfunkinhaber interviewt werden. Das Design der Erhebung ist so konstruiert, dass eine Reihe der oben genannten Probleme bearbeitet werden können.

Insgesamt bedarf die Entwicklung in der Mobilfunktelefonie der Beobachtung durch Stichprobenstatistiker. Auch die Nutzung von „Voice over IP"-Diensten wird sich auf die Ausstattung der Haushalte mit Festnetztelefonen auswirken. So

heißt es im Tätigkeitsbericht 2004/2005 der Bundesnetzagentur: „Der Wettbewerb erhält in jüngster Zeit noch zusätzliche Impulse durch neue und günstige Angebote im Mobilfunk, die die Tendenz zu einem Plattformwettbewerb zwischen Festnetz und Mobilfunk verstärken. Auch die Angebote von VoIP-Diensten zeigen in Verbindung mit der Ausbreitung der Breitbandzugänge neue Perspektiven für Innovation und Angebotsvielfalt auf."

Literatur

ADM e.V. (2005): www.adm-ev.de

Bundesnetzagentur: Tätigkeitsbericht 2004/2005 der Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen. http://www.bundesnetzagentur.de/media/archive/4515.pdf

Gabler, S./ Häder, S. (2002): Idiosyncrasies in Telephone Sampling - The Case of Germany. International Journal of Public Opinion Research. Vol. 14 No. 3: 339-345.

Müller, D. (2005): Deutsche ITK-Branche steht kerngesund da. www.zdnet.de/itmanager/kommentare/0,39023450,39137237,00.htm

Stichprobenziehung in der Umfragepraxis Die unterschiedlichen Ergebnisse von Zufallsstichproben in face-to-face-Umfragen

Jürgen H.P. Hoffmeyer-Zlotnik

Zusammenfassung

Das ADM-Master-Sample stand lange Zeit für das Design von Zufallsstichproben in face-to-face-Umfragen. Mit abnehmender Bereitschaft der Befragten, an Interviews teilzunehmen, und mit zunehmenden Kosten für die Realisierung von Interviews wurde das ADM-Master-Sample modifiziert, vereinfacht, in Aufwand und damit Kosten eingeschränkt. Nun stellt sich nach all diesen Modifizierungen die Frage nach der Qualität der Daten. Sind Veränderungen in den Aussagen der Befragten als sozialer Wandel oder als Änderung des Stichproben-Designs zu interpretieren? Wie groß sind die Effekte des Stichproben-Designs? In einem Vergleich von drei Datensätzen, die mit unterschiedlichen Stichproben-Designs durchgeführt wurden, soll diesen Fragen begegnet werden.

1 Fragestellung

In der Praxis der Sozialforschung in Deutschland kommen bei face-to-face-Erhebungen als Zufallsstichproben in erster Linie sehr unterschiedliche Varianten von Random-Walk-Stichproben zur Anwendung. Hinzu kommen auf Registern basierte Zufallsstichproben. Die Varianten der Random-Walk-Stichproben werden unter einer verwirrenden Vielzahl von Bezeichnungen angeboten, wobei ein Teil der Bezeichnungen institutsspezifisch sind und sich durchaus im Laufe der Zeit ändern können. Dieses gilt z.B. für die Bezeichnung des „Standard-Random“, wohinter sich im Laufe der Zeit unterschiedliche Varianten verbergen können, je nachdem, was von welchem Institut gerade als „Standard“ angeboten wird. Wichtig für das Verständnis ist: Die unterschiedlichen Varianten der Random-Stichproben erfordern einen unterschiedlichen Aufwand für die Rekrutierung der zu Befragenden in einem Sampling-Point, und unterschiedlich hoher Aufwand schlägt sich in unterschiedlich hohen Kosten, aber auch in einer unterschiedlichen Erreichbarkeit von spezifischen Zielpersonen und damit verbunden, in einer unterschiedlichen „Schieflage“ der Daten nieder. Das Problem für den Forscher ist darin zu sehen, dass die Geldgeber, wie z.B. die nationale Forschungsförde-

rung (ob Deutsche Forschungsgemeinschaft, DFG oder ministerielle Forschungsförderung), bei Umfragen auf die Kosten sehen und fordern, dass das im Preis günstigste Angebot den Zuschlag erhält. Während man bei Geräten problemlos über die technischen Daten eine Vergleichbarkeit der Angebote herstellen kann, ist dies bei Umfragen sehr schwierig. Es wird in der Regel nicht mit einem Standard an technischen Daten argumentiert: Während ein „ADM-Stichproben-Design" definiert ist, ist eine Stichprobe „in Anlehnung an ADM" der freien Definition der Institute ausgesetzt und damit die Bandbreite der Qualität solch eines Stichprobenplans sehr groß. Auch der Begriff des „Standard-Random" hilft hier nichts, da dieser von unterschiedlichen Instituten für unterschiedliche Stichproben-Designs benutzt wird. Mit dem Einbeziehen des Laptops in die Interviewtätigkeit hat sich die Qualität der Interviews über die elektronische Steuerung des Gesprächs und der Antwortaufzeichnung zwar verbessert, die Qualität der Stichproben ist jedoch oft schlechter geworden. Grund hierfür sind sich wandelnde Interviewerstäbe, in Abwendung vom Freizeitinterviewer und Hinwendung zum Berufsinterviewer. Schließlich müssen sich die teuren Geräte amortisieren. Dieses erfordert eine Erhöhung der Interviewzeiten pro Interviewer. Bei einer am Erfolg realisierter Interviews orientierten Bezahlung bedeutet dies, dass der Leerlauf durch das Aufsuchen von Haushalten oder Personen, die nicht leicht erreichbar und zu einem Interview bereit sind, reduziert wird. Die Konsequenz ist auch ein Wandel im Design der Stichprobenpläne.

Die überaus interessante Frage, die sich angesichts dieser Entwicklung stellt, ist: Zu welchen Ergebnisverzerrungen führt der Wandel im Design der Stichprobenpläne?

Die Frage ist nicht einfach zu beantworten, da eine große Anzahl von unterschiedlichen Einflussfaktoren auf die Qualität einer Stichprobe einwirken.

Neben dem Ziehungsdesign der Stichprobe sind dies:

- demographische und ethnische Merkmale der zu befragenden Population;
- die Nichterreichbarkeit von Zielpersonen;
- Erhebungsmethode;
- Erhebungsinstrument und
- demographische Zusammensetzung und Schulung der Interviewer.

Alle diese Faktoren beeinflussen sich auch gegenseitig. Viele dieser Faktoren werden seit Jahren untersucht:

Eine Auseinandersetzung mit dem Einfluss der Erhebungsmethode, dem Modeeffekt auf die Daten wurde durch den Einsatz von Telefon und Computer notwenig. Wie groß diese Effekte sind, zeigte sich bei der Umstellung von Zeitreihen von paper-and-pencil-Techniken zu computergestützten Techniken im face-to-

face-Bereich (Allbus 2000) und bei der Umstellung von face-to-face zu computerunterstützten Telefoninterviews (CATI) (u.a. Schulte 1997).

Mit den Bedingungen und Auswirkungen des Nonresponse (Schnell 1997) beschäftigt sich, ausgelöst über eine absinkende Teilnahmebereitschaft zum Interview und damit einhergehenden Verzerrungen bei den Antworten (Koch 1993), ein seit 1990 jährlich stattfindender Workshop on Household Nonresponse (http://www.nonresponse.org). Den Einfluss der Bedingungen im Befragtenhaushalt auf die Befragungsbereitschaft hat grundlegend für Deutschland Sodeur (1997) untersucht. Der Einfluss des Erhebungsinstrumentes wird in vielfältigen Facetten seit Jahrzehnten untersucht. Der Einfluss des Interviewerstabes wird in der so genannten „Institutshandschrift" sichtbar. Hier ist bisher nur wenig Forschung betrieben worden. Allerdings versucht man dem durchaus bekannten Phänomen durch das Einbeziehen mehrerer Institute in eine Studie (siehe Media-Analyse; vgl. TAB 2005) entgegen zu steuern. Auch die Frage nach dem Einfluss des Stichprobendesigns ist bisher wenig untersucht worden (von der Heyde 1994; Koch 1997). Die Sozial- und Marktforschungsinstitute setzen zur Korrektur dieses Einflusses die Wunderwaffe der Gewichtung ein (von der Heyde 1994; siehe auch Gabler 1994).

2 Schematische Darstellung der beiden gebräuchlichsten Zufallsauswahlen von Personen für face-to-face-Befragungen

Die gebräuchlichsten Zufallsauswahlen für face-to-face-Interviews in der Bundesrepublik Deutschland sind a) (basierend auf einem Random-Walk) das ADM-Master-Sample und b) (basierend auf den Einträgen der Einwohnermeldeämter) eine Personenstichprobe.

Das ADM-Master-Sample stellt ein dreistufiges Auswahlverfahren dar, bei dem die Auswahl der Sampling-Points nach Wahlbezirken (Flächenauswahl), die Auswahl von Haushalten in den gezogenen Sampling-Points (reine Zufallsauswahl) und die Auswahl von Zielpersonen in den gezogenen Haushalten (systematische Zufallsauswahl per Kish-Table oder durch Zufallsstart) in Stufen hintereinander geschaltet sind.

Die auf den Einwohnermelderegistern aufbauende Registerstichprobe stellt ein zweistufiges Verfahren dar, bei dem zunächst die Auswahl der Sampling-Points nach Gemeinden stattfindet und zu den Registern führt (geschichtete Zufallsauswahl), aus denen die Zielpersonen ermittelt werden (systematische Zufallsauswahl durch Zufallsstart).

Schema 1 zeigt die Stufen der Stichprobenziehung für die zwei konkurrierenden Modelle: Links sind die Stufen der Ziehung von Random-Walk-Stichproben

aufgelistet, rechts die Stufen der Ziehung einer (Einwohnermeldeamts-) registerbasierten Zufallsstichprobe.

Die erste Stufe der Stichprobenziehung stellt die Auswahl der Sampling-Points dar: Bei der Random-Walk-Stichprobe werden die Sampling-Point-Units als „Begehungseinheiten" definiert, die überschaubare und homogene Siedlungsteileinheiten darstellen sollen. Einheiten in einer Größe von 1.000 bis 2.000 Zielpersonen sind ideale Begehungseinheiten. Kleine Auswahleinheiten werden zu Begehungseinheiten von mindestens 400 potenziellen Zielpersonen aggregiert.

Bei der Registerstichprobe werden die Sampling-Point-Units über die Registereinheiten definiert. Bei der Einwohnermeldeamtsstichprobe sind dies in der Regel die Gemeinden.

Schema 1: Auswahlstufen von Zufallsstichproben über Random-Walk und Random-Register

Stufe 1: Auswahl der Sampling-Point-Units (SPU)		
	Random-Walk	Random-Register
Definition der SPU	Begehungseinheit	Registereinheit
Auswahl	als systematische Zufallsauswahl aus strukturierter Anordnung über Land, Regierungsbezirk, Kreis, Gemeinde, Regionsgrößenklasse	
Abgrenzung	Teilgemeinde	Gemeinde
Stufe 2: Auswahl der Haushalte		
Definition der Einheit	privater Haushalt	entfällt
Auswahl über ...	systematische Auflistung der Haushalte in Schrittweite nach Begehungsanweisung	entfällt
Ergebnis	Haushalt mit Adresse	entfällt
Stufe 3: Auswahl der Zielperson		
Definition der Einheit	Person mit Merkmalen	Person mit Merkmalen
Auswahl über ...	systematische Auflistung aller Personen im Haushalt und Anwendung einer Kish-Table	EDV-gestützte Zufallsauswahl aus dem Register
Ergebnis	Person in Haushalt mit Adresse	Person mit Adresse

Die Ziehung der Sampling-Points erfolgt mit Hilfe einer systematischen Zufallsauswahl aus der strukturierten Anordnung der Wahlbezirke bzw. der Gemeinden, länderweise, pro Land nach Regierungsbezirken, pro Regierungsbezirk nach

Kreisen, pro Kreis nach Regionsgrößenklassen – und bei einer Random-Register-Stichprobe pro Kreis nach Gemeinden und diese nach Regionsgrößenklasse.

Die Auswahlstufe 2 ist die Begehung für die Random-Walk-Stichprobe. Bei der registerbasierten Stichprobe entfällt eine Begehung. Die bei der Begehung zu ermittelnde Einheit ist der private Haushalt. Die Auswahl der Haushalte geschieht über eine systematische Auflistung der Haushalte, in vorgegebener Schrittweite, mittels einer Begehung nach Begehungsanweisung. Das Ergebnis ist ein Haushalt, versehen mit einer Adresse.

Zur Begehung und Auflistung von Haushalten im Sampling-Point wird dem Interviewer eine zufällig ermittelte Startadresse aus den jeweiligen Wahlbezirksunterlagen vorgegeben. Die für die Adressenauflistung relevante Begehungsanweisung legt den Weg des Interviewers, von der Startadresse ausgehend, fest. Die Route des Interviewers wird gesteuert durch Vorgaben ...

- von Straßenseiten über „gerade“ und „ungerade“ Hausnummern,
- von Richtungen über „aufsteigende“ / „fallende“ Hausnummern,
- von Richtungsänderungen über das Verhalten an Kreuzungen und
- von Möglichkeiten des Neustarts nach dem Abarbeiten von Straßenabschnitten.

Jeder als Zielhaushalt in vorgegebener Schrittweite aufgelistete Haushalt (direkte Nachbarschaft von zwei Zielhaushalten soll vermieden werden) zählt dann zum Brutto der Stichprobe. In diesen Haushalten wird zur Ermittlung jeweils einer Befragungsperson die dritte Stufe des Auswahlplans angewandt, d.h. es werden die Zielpersonen ermittelt:

Die Zielperson ist eine Person, welche einerseits die Definitionsmerkmale der Grundgesamtheit aufweist – wie z.B. Alter, ethnische Zugehörigkeit, Wahlberechtigung oder ähnliches mehr – und andererseits über eine systematische Zufallsauswahl ausgewählt wird.

Bei einer Random-Walk-Stichprobe erfolgt die Auswahl einer Zielperson über eine systematische Auflistung aller Personen im Haushalt und einer mit dieser Auflistung verbundenen Anwendung einer Wahrscheinlichkeitstabelle. Die Auswahlchance für eine Zielperson ist umgekehrt proportional zur Haushaltsgröße. Das Ergebnis ist eine Person als Repräsentant eines Haushaltes.

Schema 2: Zielpersonenbestimmung mit Kish-Table

HH	ZP	Jahrg.	Sex	a: Haushaltsgröße/b: Kennziffer								
1	1	1935	m	a:	1	2	3	4	5	6	7	8	9
	2	1941	w	b:	1	**1**	3	2	4	2	3	5	3
2	1	1920	w										
	2	1946	w	a:	1	2	3	4	5	6	7	8	9
	3	1948	m	b:	1	1	2	**3**	1	4	5	7	5
	4	1972	m										

HH = Haushalt, ZP = Zielperson, Jahrg. = Geburtsjahrgang als Sortierkriterium

Schema 2 zeigt die Anwendung der Kish-Table. Im Haushalt Nr. 1 sind 2 Personen vorhanden und nach dem Alter sortiert aufgelistet. Die Haushaltsgröße 2 in der Zeile a: verweist auf die darunter liegende Zufallszahl 1 in Zeile b. Also ist die erste Person in der Auflistung die zu befragende Zielperson. Im Vier-Personen-Haushalt Nr. 2 ist die dritte Person, identifiziert über die Zufallszahl 3 (in Zeile b:), – unter der die Haushaltsgröße angebenden Zahl 4 (in Zeile a:) – die zu befragende Zielperson.

In einem personenregisterbasierten Stichprobendesign erfolgt die Auswahl der Befragungspersonen als EDV-gestützte Zufallsauswahl aus dem Personenregister (z.B. dem Einwohnermelderegister). Das Ergebnis ist eine Person mit Adresse. Die Zwischenstufe Haushalt entfällt.

Bei einer Random-Walk-Stichprobe erfolgt noch eine Bearbeitung der Daten durch Gewichtung, da über die Zwischenstufe des Haushalts Personen aus kleinen Haushalten eine größere Chance haben, in die Stichprobe zu gelangen, als Personen aus großen Haushalten (1/n). Die notwendige Korrektur geschieht über eine Transformation per Gewichtung als fallweise Multiplikation der befragten Person mit der Anzahl der Zielpersonen im Haushalt. Das Ergebnis ist jetzt eine Person – ohne Haushaltseinbindung.

3 Varianten von Random-Walk-Designs

Das Auswahlverfahren, das auf den Einwohnermelderegistern aufbaut, kann wenig variiert werden, denn die Auswahl der Gemeinden hat nach den Regeln der Kunst geschichtet nach Lage, Größe und Gemeindetyp zu erfolgen, die Auswahl der Zielpersonen hat als systematische Zufallsauswahl durch vorgegebene Schrittweite und Zufallsstart stattzufinden. Hier können bestenfalls einzelne Gemeinden durch andere ersetzt werden, sei es aus Kostengründen oder aus man-

gelnder Kooperationsbereitschaft. Bei den Random-Route-Stichproben nach ADM-Master-Sample gibt es zwar ein vorgegebenes Prozedere, aber viele Möglichkeiten der Variationen.

Auf der Ebene der Sampling-Points sollte idealerweise kein Abweichen von den Vorgaben stattfinden. Aber alle Institute ersetzen sogenannte „problematische" Sampling-Points. Als „problematisch" akzeptiert werden oft die „Hallig" (als Synonym für Nicht-Erreichbarkeit) und die „Herbertstraße" (als Synonym für das Rotlichtmilieu).

Werden aber zu viele Sampling-Points ersetzt, oder wird „problematisch" unakzeptabel über Interviewermangel vor Ort definiert, dann erhält die Stichprobe einen Bias.

Auf der Ebene der Zielpersonenauswahl darf es keine Möglichkeit der Variation geben, da hier jedes Abweichen von der Zufallsauswahl in den Bereich der Fälschung fällt.

Auf der zweiten Stufe der Stichprobenziehung per Random-Walk gibt es viele Möglichkeiten der Variation, die den Aufwand der Identifizierung des Zielhaushaltes reduzieren und Zielhaushaltidentifikation und Zielpersoneninterview in einen Bearbeitungsschritt versetzen. Varianten der Begehung sind für den Forscher deshalb besonders tückisch, weil weder Verfahren überblickt werden noch damit verbundene Konsequenzen absehbar sind. Und von den Datenerhebungsinstituten werden Varianten der Begehung, mit klangvollen Namen versehen, als „optimale" Modifikationen angepriesen: Aus dem „Random-Route", heute „Adress-Random" genannt, wird über Zusammenlegen von Arbeitsschritten ein „Street-Random", und dieses mutiert durch ein weiteres Lockern der Vorgaben für den Interviewer zu einem „vereinfachten" Modell, das heute von einigen großen multinationalen Erhebungsinstituten zum „Standard-Random" erklärt wird.

Schema 3 zeigt die Möglichkeiten der Variationen beim Random-Walk in der zweiten Auswahlstufe, in der das Prozedere der Auswahl der Haushalte über Begehung festgelegt wird.

Schema 3: Verfahrensmodelle des Random-Walk bei der Ermittlung der Haushalte

Arbeitsschritte	**Modell A, streng**	**Modell B, mittel**	**Modell C, leicht**
Erhebung	Adressenvorlauf	Adressenermittlung integriert	
Vorgabe	Bruttovorgabe von X Adressen, Netto offen		Nettovorgabe, Brutto offen
Nachbearbeitung	zwingend	möglich	nein
Protokoll	ja	möglich	nein

Das Modell A geht von der Auflistung einer vorgegebenen Anzahl von Haushalten nach vorgegebener Begehungsroute und Schrittweite aus. Das Besondere ist

1. die separate, vom Prozess des Interviewens getrennte Begehung und Auflistung der Adressen durch Interviewer X zum Zeitpunkt t1;
2. eine von der Adressenauflistung getrennte systematische Zufallsauswahl der Zielhaushalte, durchgeführt im Erhebungsinstitut zum Zeitpunkt t2;
3. die in einem dritten Schritt folgende Zielpersonenauswahl per Kish-Table und das Durchführen der Interviews. Dieses wird vom Interviewer Y zu einem Zeitpunkt t3 durchgeführt.
   Die Vorgabe für den Interviewer lautet: In jedem der vorgegebenen Haushalte eine Zielperson auszuwählen und möglichst auch zu interviewen.
4. Erreicht Interviewer Y eine zu geringe Anzahl an realisierten Interviews, so wird Interviewer Z zum Zeitpunkt t4 zu einer Nachbearbeitung geschickt, soweit dies die Bedingungen des Datenschutzes erlauben, Haushaltkontaktperson oder Zielperson noch nicht definitiv verweigert haben.

Beim Modell B listet Interviewer X den Haushalt in vorgegebener Schrittweite auf, ermittelt im gleichen Arbeitsschritt die Zielperson und führt das Interview durch. Alles geschieht in einem und demselben Arbeitsschritt. Aufgelistet wird bei der Begehung ein Brutto von N Haushalten. Aus diesem Brutto der N Haushalte sind so viele Interviews wie möglich zu realisieren. Eine Nachbearbeitung ist nicht zwingend erforderlich. Das Modell B war lange Zeit das gebräuchlichste Verfahren. Die Kosten sind gegenüber dem Modell A etwa 20 % niedriger. Aber die Kosten des Modells B lassen sich noch einmal um etwa 20 % reduzieren, wenn die Bruttovorgabe von Adressen, unter denen ein Maximum an Interviews zu realisieren ist, durch eine Nettovorgabe von zu realisierenden Interviews ersetzt wird. Die Nettovorgabe verzichtet auf die Festsetzung eines Brutto von maximal erlaubten Adressen.

Das Modell C stellt jenes Stichprobendesign dar, das einige Institute als „Standard“ durchsetzen wollen. Die Regieanweisung lautet wie beim Modell B. Der Unterschied besteht darin, dass Interviewer X nicht mehr durch eine als Brutto vorgegebene Anzahl von Haushalten beim Interviewen eingeschränkt wird, sondern so viele Haushalte auflisten und kontaktieren kann, wie er benötigt, um Netto = N Interviews zu realisieren. Durch die fast beliebige Verlängerung der Begehung wird eine Nachbearbeitung überflüssig. Eine Dokumentation der Kontakte findet nicht statt, da ein kontrollierter Adressenrahmen fehlt.

4 Vergleich der Güte der Realitätsabbildungen dreier Stichprobendesigns mit dem Mikrozensus

Für den nachfolgenden Vergleich von drei Datensätzen, die nach drei unterschiedlichen Stichprobendesigns erhoben wurden, werden folgende Datensätze benutzt:

- für das strenge Modell des Random-Walk: Allbus 1998
- für das leichte Modell des Random-Walk: Allbus 1992
- für eine Registerstichprobe aus dem Einwohnermeldeamt: Allbus 1996

In die Analyse werden nur die Daten aus den alten Bundesländern einbezogen.

Da die drei verglichenen Stichproben unterschiedliche Designs darstellen, sind unterschiedlich starke Effekte eines Abweichens vom Referenzwert zu vermuten. Die Referenzstatistik stellt der Mikrozensus, die von der amtlichen Statistik jährlich erhobene Ein-Prozent-Stichprobe der Wohnbevölkerung, dar.

- Das strenge Modell des Random-Walk bietet wenig Spielraum zur Improvisation, da mehrere Interviewer pro Fall eingesetzt werden und hohe Kontrollen den Random-Walk transparent machen. Die Fehlerquote des Random-Walk ist gering und in der Regel nachvollziehbar.
- Das leichte Modell des Random-Walk plädiert an die Ehrlichkeit der Interviewer, weil es kaum kontrollierbar ist. Es ermöglicht dem Interviewer, sich auf die Leicht-Erreichbaren zu beschränken.
- Die Registerstichprobe schränkt die Möglichkeit zur Manipulation maximal ein, da eine konkrete Person mit Name, Vorname, Straße, Hausnummer vorgegeben ist. Diese Person nicht zu kontaktieren bedarf einer guten Begründung.

Die Ausgangshypothesen für den Stichprobenvergleich heißen:

- Je effektiver der Stichprobenplan dem Interviewer die Möglichkeit zu einer vom Plan abweichenden Zielpersonenauswahl nimmt, desto besser die Annäherung an die Zufallsstichprobe.
- Je mehr Möglichkeiten der Interviewer hat, unkontrolliert Einfluss auf die Auswahl der Zielpersonen auszuüben, desto größer die Möglichkeit einer systematischen Abweichung von einer „sauberen“ Zufallsstichprobe.
- Der Mikrozensus kommt dem „wahren“, aber dennoch unbekannten Wert der Grundgesamtheit am nächsten.

Tabelle 1 zeigt, eingeschränkt auf die Befragten in den alten Bundesländern, die Verteilungen der vier Variablen: „Geschlecht“, „Alter“, „Bildung“ und „Zugehörigkeit zu einem Ein-Personen-Haushalt“ in den drei Stichproben-Datensätzen und dem Mikrozensus von 1997:

Spalte 1 beinhaltet das Modell C des Random-Walk, die Netto-Stichprobe, Spalte 2 das kontrollierte Random-Walk mit Vorauflistung der Adressen und einer Brutto-Vorgabe an zu realisierenden Fallzahlen, Spalte 3 die Registerstichprobe.

Geht man davon aus, dass der Mikrozensus die Abbildung bietet, die die Grundgesamtheit am fehlerfreiesten beschreibt, und geht man weiter davon aus, dass der soziale Wandel – diese benutzten vier Variablen betreffend – in den alten Bundesländern im Zeitraum von 1992 bis 1998 nicht ins Gewicht fällt, so ist der Grad der Übereinstimmung mit der Mikrozensusverteilung als Gütekriterium für die Umfragedaten zu sehen. Daher sind in Spalte 4 zum Vergleich die Daten des Mikrozensus 1997 für die alten Bundesländer ausgewiesen.

Tabelle 1: Abweichungen der Verteilungen demographischer Variablen der einzelnen Stichprobendesigns vom MZ 1997

	Modell C, leicht **Allbus 1992**	Modell A, schwer **Allbus 1998**	Register **Allbus 1996**	MZ 97
Alter				
18 – 29	+4,2	-0,1	+3,8	17,1
30 – 39	+1,6	+0,1	+1,8	19,7
40 – 49	+0,7	-0,1	+0,8	16,8
50 – 59	+0,6	+1,8	+0,5	17,0
60 – 69	-1,6	+2,3	-1,3	14,7
70 plus	-5,8	-3,9	-5,5	14,6
Bildung				
9. Klasse	-3,7	-6,9	-7,1	56,8
10. Klasse	+1,9	+5,0	+2,4	23,3
12./13. Kl.	+1,8	+1,7	+4,6	19,9
Haushaltsgröße				
1-Pers-HH	-8,3	-5,5	-3,6	20,7

Bei der Variablen „Alter“ zeigt sich ein Institutseffekt: Die Erhebungen von 1992 und von 1996, also das am wenigsten kontrollierte Design der Netto-Stichprobe und das am restriktivsten kontrollierte Design der Einwohnermeldeamtsstichprobe, sind vom Institut A durchgeführt, die Erhebung von 1998, das kontrollierte Random-Walk, von Institut B. Davon ausgehend, dass das Maß der Kontrolle einen Einfluss auf die Güte der Daten hat, müssten sich die Netto-Stichprobe und die Register-Stichprobe am stärksten unterscheiden. Bei der Variablen „Alter“ zeigen die Netto-Stichprobe und die Register-Stichprobe eine sehr ähnliche Ver-

teilung: In beiden Stichproben sind die jungen Zielpersonen der Altersgruppe der 18- bis 29-jährigen mit etwa + 4 % überrepräsentiert und die alten Zielpersonen der Altersgruppe „70 und älter“ mit über - 5 % unterrepräsentiert. Beim kontrollierten Random-Walk zeigt sich ein gegenläufiger Effekt: Die jungen Altersgruppen der 18- bis 49-jährigen entsprechen genau der Verteilung im Mikrozensus (Abweichungen um 0,1 %), die Altersgruppen der 50- bis 69-jährigen sind um 2 % überrepräsentiert und die älteste Gruppe der über 69-jährigen ist deutlich geringer unterrepräsentiert als bei den anderen beiden Stichproben-Designs. Dieser Effekt hat nichts mit einem Effekt des Stichproben-Designs zu tun, sondern lässt auf eine institutsspezifisch unterschiedliche Schulung und Instruierung der Interviewer schließen: Die Institute definieren die Problemgruppen, auf die besonders zu achten sind, unterschiedlich.

Bei der Variablen „Bildung“ zeigt sich in erster Linie ein Interviewereffekt: die niedrig Gebildeten sind durchgängig unterrepräsentiert und die höher Gebildeten sind über alle drei Studien überrepräsentiert. Grund hierfür ist der hohe Anteil der höher Gebildeten unter den Interviewern (bei der Umfrage 1996 wiesen nur etwa 25 % der Interviewer einen niedrigen Bildungsabschluss auf) und die bessere Kontaktmöglichkeit unter gleichem Bildungsniveau. Bei der Variablen „Bildung“ zeigt sich allerdings auch ein leichter Design-Effekt: Beim nicht kontrollierten Random-Walk mit Netto-Vorgabe, dem Design, das ein Über-Repräsentieren der Leicht-Erreichbaren ermöglicht, sind die Zielpersonen der unteren Bildungsgruppe deutlich weniger überrepräsentiert (- 3,7 %) als bei den stark kontrollierten Stichproben-Designs. Hier macht sich bemerkbar, dass die niedriger Gebildeten und Nicht-Erwerbstätigen leichter zu erreichen sind als andere Gruppen.

Der Designeffekt kommt erst richtig zum Tragen bei der Variablen „Ein-Personen-Haushalte“. Ein-Personen-Haushalte stellen eine zentrale Gruppe der Schwer-Erreichbaren dar. Tabelle 1 zeigt eindrucksvoll, wie der Anteil der Befragten aus Ein-Personen-Haushalten mit zunehmender Restriktion bei der Zielpersonenermittlung zunimmt.

5 Kontrolle der Qualität unterschiedlicher Stichproben-Designs durch einen Vergleich von Paaren

Um Effekte unterschiedlicher Stichprobendesigns sichtbar zu machen, darf man sich nicht blind auf die Daten verlassen, die die Institute ausliefern. Institute können die Daten vor der Auslieferung per Datenbereinigung und/oder Gewichtung an eine gewünschte Realität oder an den aktuellen Mikrozensus anpassen.

Um die Effekte sichtbar zu machen, muss man sich auf Merkmale konzentrieren, deren Verteilung bei einer unverzerrten Zufallsauswahl feststehen. Da bei einer unverzerrten Zufallsauswahl jede Person in einem Zwei-Personen-Haushalt

eine etwa gleiche Auswahlchance von ca. 50 % haben muss, bieten sich Paare als ideale Analyseeinheit an. In dem nun folgenden zweiten Analyseschritt werden nur noch jene Zielpersonen betrachtet, die in Paarbeziehungen in einem Haushalt leben, der exakt zwei zur Grundgesamtheit zählende gegengeschlechtliche Personen – d.h. ein volljähriges Paar mit einer männlichen und einer weiblichen Person, die in einer partnerschaftlichen Beziehung zueinander stehen – aufweist. Minderjährige Kinder bleiben als nicht zur Grundgesamtheit zählend unberücksichtigt. Damit gehen in die nun folgende Analyse alle zusammenlebenden gegengeschlechtlichen Paare ein, egal, ob minderjährige Kinder im Haushalt leben oder nicht (siehe Sodeur 1997). Abweichungen von der 50 zu 50 Verteilung des Geschlechts bei gegengeschlechtlichen Paaren sind entweder durch das Verhalten der Zielpersonen (Verweigerungen oder Abwesenheiten) oder durch das Verhalten der Interviewer (zu wenig Kontaktversuche, Konzentration auf die Leicht-Erreichbaren, Abweichen von der Kish-Table) verursacht. Das Verhalten der Zielpersonen dürfte sich über die einzelnen Studien von 1992 bis 1998 – zumindest in den alten Bundesländern – nicht wesentlich verändert haben. Das Verhalten der Interviewer ist abhängig vom Stichprobendesign und den ihnen auferlegten Restriktionen bei der Zielpersonenauswahl.

Tabelle 2 zeigt zwischen den einzelnen Studien leichte Unterschiede beim Anteil der „Erwerbstätigen Frauen“, beim Anteil der „Geschlechter“ und beim Anteil der „Frauen je Altersgruppe“. Diese Unterschiede lassen sich auf die unterschiedlichen Stichprobendesigns zurückführen: Die Stichprobendesigns unterscheiden sich im Erreichen der Zielpersonen!

Die Erreichbarkeit der vollzeit-erwerbstätigen Frau ist Design-bedingt unterschiedlich:

- Die vollzeit-erwerbstätige Frau ist schwerer zu erreichen als diejenige, die dieses Merkmal nicht aufweist. Bei dem Stichprobendesign, das dem Interviewer am wenigsten Restriktionen auferlegt, mit dem leichten Modell C, werden die wenigsten vollzeit-erwerbstätigen Frauen erreicht (21 %). Bei dem Stichprobendesign, welches über Name und Adresse der zu interviewenden Person die Zielperson schon vor dem Erstkontakt mit dem Haushalt klar definiert hat und darüber hinaus den Interviewer maximal kontrollieren kann, wird der höchste Anteil an erwerbstätigen Frauen erreicht (27 %).

Die Erreichbarkeit von Männern und Frauen widerspricht den Erwartungen nur auf den ersten Blick:

- Bei den Paaren ohne minderjährige Kinder überwiegen die Interviews mit den Männern. Diese Befragtengruppe umfasst jedoch einen hohen Anteil an alten Personen – eine Gruppe, in der die Männer erreichbar sind und die Rolle der Außendarstellung des Haushalts pflegen. Daher drängen sich ältere, in der Wohnung anwesende Männer zum Interview. Aus Erfahrung wissen wir, dass

in solchen Fällen der Interviewer Schwierigkeiten hat, die nach dem Kish-Table ausgewählte Frau wirklich interviewen zu können.

Tabelle 2: Verteilungen demographischer Variablen von Befragten unterschiedlicher Stichproben, die mit ihrem/r Partner/in in einem Haushalt von 2 Personen im Alter ab 18 Jahren zusammenleben

	Modell C, leicht **Allbus 1992**		Modell A, schwer **Allbus 1998**		Register **Allbus 1996**	
	Anteil	Abweichung	Anteil	Abweichung	Anteil	Abweichung
Erwerbstätige Frau						
vollzeit erwerbstätig	21,2		25,5		26,9	
N	751		656		778	
Paar ohne Kind						
Frau	43,6	-6,4	46,0	-4,0	50,8	+0,8
N	870		739		762	
Paar mit Kind						
Frau	53,0	+3,0	53,7	+3,7	45,3	-4,7
N	940		739		864	
Anteil Frau je Altersgruppe						
20 - 29	63,8	+13,8	61,9	+11,9	54,5	+4,5
30 – 39	58,7	+8,7	59,0	+9,0	51,0	+1,0
40 – 49	56,2	+6,2	46,3	-3,7	48,0	-2,0
50 – 59	44,8	-5,2	49,1	-0,9	45,3	-4,7
60 – 69	31,6	-18,4	40,5	-9,5	41,9	-8,1
70 +	26,7	-23,3	37,4	-12,6	43,7	-6,3
ungewichtet						

- Bestätigung findet dieser Befund bei einem Blick auf den Frauenanteil der Interviewten in den Altersgruppen ab 50 Jahre. Auch hier kommt das Stichprobendesign zum Tragen, das beim Modell C, der Netto-Stichprobe, dem Interviewer die Möglichkeit gibt, dem leichteren Kontakt nachzugeben.
- Bei Interviews mit Paaren mit Kindern zeigen sich zwischen Random-Walk-Verfahren und Register-Stichprobe entgegengesetzte Übersteuerungen. Beides hat wieder mit Erreichbarkeit zu tun: Ein Random-Walk-Verfahren trifft bei der Begehung auf die häufiger in ihrer Wohnung anwesende und damit leichter erreichbare Frau. Die Register-Stichprobe gibt Personen vor, die vor

dem Interview zwecks Terminvereinbarung telefonisch kontaktiert werden können. Ein Interview von 50 Minuten Länge stellt für eine Mutter mit Kind oft eine Störung dar und lässt sich beim telefonischen Vorabkontakt einfacher ablehnen als an der Wohnungstür. Die weniger gestörte Verfügbarkeit für ein Interview weist in der Familie mit Kind oft der Mann auf.

- Ein Blick auf den Anteil der befragten Frauen pro Altersgruppe zeigt, dass mit abnehmender Restriktion im Stichprobendesign (von Modell C zu Register) der Anteil der jungen Frauen unter den Befragten zunimmt – der Anteil der 20- bis 29-jährigen ist besonders hoch bei der Netto-Stichprobe (Modell C) – während der Anteil der älteren Frauen abnimmt – der Anteil der ab 60-jährigen ist besonders niedrig bei der Netto-Stichprobe (Modell C). Hier haben wir wieder einen eindeutigen Einfluss des Stichprobendesigns, denn die älteren Frauen haben eine größere Angst, den fremden Interviewer in die Wohnung zu lassen, und gehören damit in die Gruppe der Schwer-Erreichbaren. In den Stichproben-Designs mit geringeren Kontrollen/Kontrollmöglichkeiten sind die Leichter-Erreichbaren in der Regel überrepräsentiert.
- Die Registerstichprobe zeigt über alle Altersgruppen eine gute Anpassung der befragten Frauen an die optimale Verteilung. Selbst die Problemgruppen der alten Frauen weichen deutlich unter 10 % von der idealen Verteilung ab. Auch findet sich in der jüngsten Altersgruppe, bei den Leicht-Erreichbaren, nur eine Übersteuerung von 4,5 %. Damit zeigt sich, dass die Registerstichprobe hinsichtlich der Merkmalskombinationen Geschlecht und Alter die beste Anpassung an eine ideale Verteilung bietet.

6 Auswirkungen der Verzerrungen durch das Stichprobendesign auf Einstellungsvariablen

Abschließend soll der Frage nachgegangen werden, in welchem Ausmaß sich die Ungleichverteilung der interviewten Personen auf die Einstellungsmessung niederschlägt. Hierzu wird der „Ist"-Anteil der befragten Frauen im Stichproben-Design Modell C (Allbus 1992) mit dem „Soll"-Anteil der Frauen pro Altersgruppe in der idealen Verteilung über ein „Soll-zu-Ist"-Gewicht verglichen. Dieses Gewicht ermittelt die Relation zwischen dem gemessenen Anteil (Ist) und dem in Paaren idealtypisch bei 50 % angenommenen Anteil (Soll). Das Ziel ist das Sichtbarmachen der Verzerrungen zwischen „Ist" und „Soll" und deren Einfluss auf die abgefragten Einstellungen.

Für die Analyse (siehe Tabelle 3) wurden willkürlich zwei Items ausgewählt:

- Item 1:
  Fragetext: „Es ist für alle Beteiligten viel besser, wenn der Mann voll im Berufsleben steht und die Frau zu Hause bleibt und sich um den Haushalt und die Kinder kümmert."
  Antwort: „stimme voll und ganz zu"
- Item 2:
  Fragetext: „Gibt es eigentlich hier in der unmittelbaren Nähe – ich meine so im Umkreis von einem Kilometer – irgendeine Gegend, wo Sie nachts nicht alleine gehen möchten?"
  Antwort: „Ja, gibt es."

Tabelle 3: Ist-zu-Soll-Gewichtung bei zwei Einstellungs-Items

Item 1:
„Es ist für alle Beteiligten viel besser, wenn der Mann voll im Berufsleben steht und die Frau zu Hause bleibt und sich um den Haushalt und die Kinder kümmert." Antwort: „stimme voll und ganz zu"

Altersgr.	20-29	30-39	40-49	50-59	60-69	70+	N
ungew.	11,7	21,7	18,9	25,0	14,4	8,3	180
gewichtet	9,2	18,5	16,8	27,9	22,8	15,5	
Differenz	2,5	3,2	2,1	2,9	8,4	7,2	

Item 2:
„Gibt es eigentlich hier in der unmittelbaren Nähe – ich meine so im Umkreis von einem Kilometer – irgendeine Gegend, wo Sie nachts nicht alleine gehen möchten?." Antwort: „Ja, gibt es"

Altersgr.	20-29	30-39	40-49	50-59	60-69	70+	N
ungew.	20,1	29,2	20,6	16,9	9,1	4,0	373
gewichtet	15,7	24,8	18,3	18,9	14,4	7,5	
Differenz	4,4	4,4	2,3	2,0	5,3	3,5	
Gewicht	0,784	0,852	0,890	1,116	1,582	1,873	

Daten: Allbus 1992

Das verwendete Gewicht ist aus der Relation „Ist" (Anteil der realisierten weiblichen Befragten) zu „Soll" (Anteil der idealerweise zu realisierenden weiblichen Befragten im Umfang von 50 %) berechnet und in der letzten Zeile von Tabelle 3 wiedergegeben.

Die weiblichen Befragten in den Altersgruppen ab 50 sind in der Stichprobe des Allbus 1992 unterrepräsentiert (siehe Tabelle 2). Die Befragten der unterrepräsentiert besetzten Altersgruppen haben aber zur „Berufstätigkeit der Frau, wenn Kinder vorhanden sind“ (Item 1) eine klare Einstellung. Ohne die vorgenommene Gewichtung wäre deren Einstellung nicht angemessen berücksichtigt worden. Im Gegensatz dazu wird ungewichtet die Einstellung der jungen Befragten, die weniger deutlich mit extremer Zustimmung antworten, zu hoch und damit verzerrend bewertet.

Ähnlich sieht es mit den Antworten auf Item 2: „Angst, nachts alleine draußen zu gehen“ aus. Hier sind die Differenzen zwischen den jungen und den älteren befragten Frauen nicht ganz so hoch. Das hier sichtbar werdende Problem ist allerdings das der Gewichtung zur Verbesserung der Einstellungsmessung. Item 2 fragt Ängste ab. Aber diejenigen, die vor unbekannten Kontakten Angst haben, stellen einen größeren Anteil des Nonresponse in den höheren Altersgruppen dar. Deren Antwortverhalten ist per Gewichtung wohl kaum zu generieren.

7 Schlussbemerkung

Die Qualität unterschiedlicher Stichproben-Designs rückte erst in den Blick, als die Restriktionen für die face-to-face-Interviewer stark gelockert wurden und das Stichproben-Design des Random-Walk mit Netto-Vorgabe den Einzug in die Umfragepraxis hielt. Allerdings orientieren sich die bisherigen Qualitätsabschätzungen an dem Abgleich der Umfragedaten mit den Mikrozensusdaten. Wie Tabelle 1 gezeigt hat, werden bei diesem Abgleich neben Designeffekten vor allem Institutseffekte und Interviewereffekte sichtbar. Und der hier sichtbar werdende Designeffekt taucht bei einer Variablen auf, die nicht zur Anpassung der erhobenen Daten an den Mikrozensus benutzt wird.

Um den Effekt des Stichprobendesigns, der von anderen Effekten stark überlagert wird, deutlich herauszuarbeiten, müssen die anderen Effekte für die Analyse eliminiert werden, was über einen „Paar-Vergleich“ möglich wird. Im „Paar-Vergleich“ zeigt sich, dass der Designeffekt durchaus nicht zu vernachlässigen ist (Tabelle 2). Berücksichtigt man bei Einstellungsdaten die im Paar-Vergleich aufgezeigten Verzerrungen zwischen realisierter und idealer Stichprobe, dann zeigt sich, dass der Designeffekt die Einstellungsdaten in beträchtlichem Maße verfälschen kann (Tabelle 3). Änderungen im Stichprobendesign können in den Daten deutlichere Änderungen als der soziale Wandel bewirken. Damit ergibt sich das Problem, dass ohne eine Berücksichtigung des Stichprobendesigns methodische Artefakte als sozialer Wandel interpretiert werden können.

Allerdings gibt es noch das Problem der Korrektur durch Gewichtung. Oder anders ausgedrückt ...

Alte Damen haben Angst, einen Interviewer in die Wohnung zu lassen. Auf die Frage nach der Angst nachts, draußen („gibt es hier eine Gegend, wo Sie nachts nicht alleine gehen möchten?“) antworten die mutigsten 54 % der Zielpersonen. Deren Verteilung wird jetzt über die „Paarwahrscheinlichkeit“ auf 100 % mit Faktor 1,873 gewichtet. Bei einer Anpassung an den Mikrozensus kommt nur der Faktor 1,06 zum Tragen. Muss aber bei den Ängstlichen nicht ein Gewichtungsfaktor angelegt werden, der deutlich über Faktor 2 liegt, da die Ängstlichen auch mehr Angst äußern würden als die weniger Ängstlichen, die sich befragen lassen? Die Anpassung an den Mikrozensus hilft bei Einstellungs-Items wenig. Die Anpassung an die Paarwahrscheinlichkeit bringt realistischere Ergebnisse, da diese die realen Gruppengrößen besser berücksichtigt. Allerdings haben wir nur Referenzen für demographische Daten und nicht für Einstellungen.

Auch beim Allbus sind manche Veränderungen zwischen den Erhebungszeitpunkten eher auf methodische Artefakte als auf sozialen Wandel zurückzuführen. Aber der Allbus ist auch eine Methodenstudie. Und ohne den Wechsel von Stichprobendesign und Erhebungsinstitut wären wir in der Abschätzung der Effekte von Stichprobendesigns deutlich weniger weit.

Literaturangaben

Gabler, S. (1994): Eine allgemeine Formel zur Anpassung von Randtabellen. In: Gabler, S. & Hoffmeyer-Zlotnik, J. H.P. (Hrsg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 88-105.

Koch, A. (1993): Sozialer Wandel als Artefact unterschiedlicher Ausschöpfung? Zum Einfluß von Veränderungen der Ausschöpfungsquote auf die Zeitreihen des ALLBUS. In: ZUMA-Nachrichten 33, S. 83-113.

Koch, A. (1997): ADM-Design und Einwohnermelderegister-Stichprobe. Stichprobenverfahren bei mündlichen Bevölkerungsumfragen. In: Gabler, S. & Hoffmeyer-Zlotnik, J. H.P. (Hrsg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 99-116.

Schnell, R. (1997): Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklungen und Ursachen. Opladen: Leske & Budrich.

Schulte, W. (1997): Telefon- und Face-to-Face-Umfragen und ihre Stichproben. Allgemeine Bevölkerungsumfragen in Deutschland. In: Gabler, S. & Hoffmeyer-Zlotnik, J. H.P. (Hrsg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag. S. 148-195.

Sodeur, W. (1997): Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. In: ZA Information 41, S. 58-82

TAB 2005: http://www.tab.fzk.de/de/projekt/zusammenfassung/ab54.htm (28-11-05).

von der Heyde, C. (1994): Gewichtung am Beispiel: Einwohnermeldeamt versus Random Route. In: Gabler, S., Hoffmeyer-Zlotnik, J. H.P. & Krebs, D. (Hrsg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 141-151.

Datensätze:

Allbus 1992
Allbus 1996
Allbus 1998

Datenqualität durch Stichprobenverfahren bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften – ALLBUS

Michael Blohm

1 Einleitung

Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) folgt dem Konzept der sogenannten ‚General Social Surveys' (vgl. Davis et al. 1994). Diese langfristig angelegten nationalen Umfrageprogramme erheben regelmäßig, unabhängig und akademisch kontrolliert, Daten zu Einstellungen, Verhalten und Sozialstruktur, um wesentliche Aspekte des gesellschaftlichen Wandels mit den Mitteln der Umfrageforschung zu erfassen. Die Daten werden der gesamten sozialwissenschaftlichen Profession für Forschung und Lehre zur Verfügung gestellt.

Seinem hohen Anspruch entsprechend bemüht sich der ALLBUS um Qualität in jeder Hinsicht. Was den Inhalt der Fragebögen angeht, um theoretisch relevante Indikatoren zu einem breiten Themenspektrum, was die Zugänglichkeit der Daten angeht, um schnelle, umfassende und nutzerfreundliche Aufbereitung und Dokumentation der Daten, und natürlich nicht zuletzt um eine hohe Datenqualität durch Minimierung des Messfehlers, wobei es etwa systematische Antwortverzerrungen durch Mode-, Kontext- oder Interviewereffekte ebenso zu minimieren gilt wie Nonresponse. Dies erfordert den Einsatz methodisch anspruchsvoller und damit häufig leider auch teurer Stichprobenverfahren und deren den Verfahrensregeln entsprechenden Umsetzung in die Praxis, was letztendlich nur mit Hilfe eines intensiven Monitorings und einer umfassenden Kontrolle der einzelnen Prozesse – einigermaßen – sicherzustellen ist.

In dem vorliegenden Beitrag wird überprüft, ob der ALLBUS durch die Wahl des Registerstichprobenverfahrens seinem Qualitätsanspruch gerecht wird. Hierzu werden die ALLBUS Nettostichproben ab 1992 mit Stichproben anderer sozialwissenschaftlicher Studien und dem Mikrozensus verglichen. Dabei steht die Frage nach dem Zusammenhang zwischen verwendeten Stichprobenverfahren, Ausschöpfungsquote und Datenqualität im Vordergrund.

2 Der Zusammenhang zwischen Stichprobenverfahren, Ausschöpfungsquote und Datenqualität

In Deutschland wird für die meisten sozialwissenschaftlichen Umfragen auf das Stichprobendesign des Arbeitskreises Deutscher Marktforschungsinstitute (ADM-Stichprobensystem) zurückgegriffen. Renommierte sozialwissenschaftliche Umfragen wie z. B. der ALLBUS und die deutsche Teilstichprobe des European Social Survey (ESS) werden hingegen unter Verwendung von Einwohnermeldeamtsstichproben durchgeführt. Diese Stichprobenverfahren haben – zumindest für Bevölkerungsumfragen – den besten methodischen Ruf, da sie die größtmögliche Kontrollsicherheit, Dokumentierbarkeit sowie Regelgebundenheit aufweisen (von der Heyde 1999: 120). Insgesamt wird den Registerstichproben hierdurch eine höhere Datenqualität nachgesagt. Den Vorteilen stehen deutlich höhere Kosten, die zum einen auf die Stichprobenbildung, insbesondere aber auf die Feldarbeit zurückgeführt werden können, gegenüber.

In einem ersten Schritt soll der Frage nachgegangen werden, ob Einwohnermeldeamtsstichproben im Vergleich zu anderen kostengünstigeren Stichprobenverfahren tatsächlich eine höhere Datenqualität aufweisen. Dies wird oft per se angenommen und auch in wenigen Arbeiten (z.B. Alt, Bien & Krebs 1991, Häder & Häder 1997) mehr oder weniger gezeigt. Leider beziehen diese Arbeiten meist nur wenige Stichproben in die Untersuchung ein, deshalb soll im Folgenden anhand von 13 bundesweiten Umfragen – durch die getrennte Analyse für die neuen und alten Bundesländern insgesamt 26 Stichproben – die Datenqualität von Stichproben in Abhängigkeit des verwendeten Stichprobenverfahrens analysiert werden.

In einem zweiten Schritt wird untersucht, ob sich ein positiver Zusammenhang zwischen der Datenqualität und den Ausschöpfungsquoten dieser Studien finden lässt. Ein solcher wird häufig postuliert. Obwohl eine Betrachtung der Formel zur Berechnung des Non-Response Bias zeigt, dass eine höhere Ausschöpfung höchstens die Wahrscheinlichkeit für einen niedrigeren Non-Response Bias erhöht, wird die Ausschöpfung oft als einziges Qualitätskriterium zur Beurteilung einer Studie herangezogen. Dass dies nicht angemessen ist, wird in vielen Studien gezeigt, die keinen Zusammenhang von Ausschöpfungsquoten und Datenqualität feststellen können (vgl. Koch 1998, Keeter et al. 2000, Curtin et al. 2000, Merkle & Edelmann 2002).

2.1 Die untersuchten Umfragen

Für die Analyse wurden Bevölkerungsumfragen ausgewählt, die alle als persönlich-mündliche Interviews durchgeführt wurden, die vergleichbare Grundgesamtheiten[1] besitzen, die annähernd die gleiche Nettofallzahlen[2] aufweisen und die für sich in Anspruch nehmen, hohen Qualitätsstandards zu genügen. Ausgangspunkt dieser Analyse ist die Arbeit von Koch (1998) in der 6 Umfragen hinsichtlich der Ausschöpfungsquoten und Stichprobenverzerrungen untersucht wurden. In dem vorliegenden Beitrag werden die von Koch (1998) untersuchten Umfragen um den ALLBUS 2000, 2002, 2004 und den deutschen Teil des European Social Survey der Jahr 2002 (Round 1) und 2004 (Round 2) ergänzt. Eine Zusammenstellung der untersuchten Umfragen findet sich in Tabelle 1, dabei sind die Umfragen nach den drei in Deutschland verbreitetsten bevölkerungsrepräsentativen Auswahlverfahren für face-to-face Umfragen gruppiert.

Tabelle 1: Übersicht über die untersuchten Umfragen

Stichprobenverfahren	Umfragen	Referenz
Random-Route	ALLBUS 1992	MZ 1993
	Wohlfahrtssurvey 1993	MZ 1993
	SowiBus II 1993	MZ 1993
	SowiBus III 1993	MZ 1993
Adress-Random	ALLBUS 1998	MZ 1997
	Media-Analyse 1994	MZ 1993
Einwohnermeldeamt (Register)	ALLBUS 1994	MZ 1993
	ALLBUS 1996	MZ 1995
	ALLBUS 2000	MZ 1999
	ALLBUS 2002	MZ 2001
	ALLBUS 2004	MZ 2003
	ESS 2002 (Dt. Teilstichprobe)	MZ 2001
	ESS 2004 (Dt. Teilstichprobe)	MZ 2003

1 Alle Umfragen beziehen sich auf die erwachsene Bevölkerung in Privathaushalten, mit Ausnahme der Media-Analyse (ab 14 Jahren) und der deutschen Teilstichproben des ESS (ab 15 Jahre). In allen Umfragen gehören Deutsche und Ausländer der Grundgesamtheit an, außer dem Wohlfahrtssurvey 93, SoWi-BUS II, SoWi-Bus III und der Media-Analyse. Diese beziehen sich nur auf die deutsche Bevölkerung.

2 Die Netto-Fallzahlen der Umfragen bewegen sich zwischen N = 2820 und N = 4166. Die Ausnahme bildet die Media-Analyse N = 25621.

Insgesamt sechs der Studien wurden nach dem dreistufigen Stichprobendesign des Arbeitskreises Deutscher Marktforschungsinstitute (ADM-Stichprobensystem) durchgeführt, wobei bei vier der Umfragen die Variante Random-Route und in zwei die Variante Adress-Random zum Einsatz kam. Bei allen ADM-Designs wird auf der ersten Auswahlstufe eine Stichprobe von Wahlkreisen gezogen. Auf der zweiten Ebene werden anhand von Begehungsregeln Haushalte und auf der dritten Ebene die Zielpersonen mit Hilfe eines Zufallschemas (z.B. Schwedenschlüssel) vom Interviewer ausgewählt. Der Unterschied zwischen der Variante Random-Route und der Variante Adress-Random besteht darin, dass bei Adress-Random die Begehung und die Befragung in getrennten Schritten erfolgt. Während bei Random-Route der Interviewer die Haushalte und die Zielpersonen in einem Schritt – anhand von Begehungsregeln – auswählt, werden bei Adress-Random zunächst die Haushalte – anhand von Begehungsregeln – gelistet. Aus diesen wird im Institut eine Stichprobe von Haushalten gezogen und an den durchführenden Interviewer weitergegeben. Dieser wählt somit „nur“ die Zielperson aus dem Haushalt aus.[3]

Die untersuchten Einwohnermeldeamtsstichproben weisen ein zweistufiges Verfahren auf. Auf der ersten Ziehungsstufe wird zunächst eine Stichprobe von Gemeinden gebildet.[4] Aus den Einwohnerregistern dieser Gemeinden werden dann die Zielpersonen für die Befragung ausgewählt. Diese werden mit Name und Anschrift den Interviewern vorgegeben.[5]

Auf einen wichtigen Unterschied der Register- und der ADM-Stichproben soll an dieser Stelle hingewiesen werden. Während bei Registerstichproben die Zielpersonen die gleiche Auswahlwahrscheinlichkeit aufweisen, gilt dies bei ADM-Stichproben nur für die Haushalte. Für Untersuchungen auf der Ebene von Personen müssen Analysen von Daten, die mit Hilfe von ADM-Stichproben erhoben wurden, transformationsgewichtet durchgeführt werden.

Theoretisch führen alle drei Verfahren zu bevölkerungsrepräsentativen Zufallsstichproben. In der Praxis der Umsetzung der Verfahren lassen sich aber Defizite identifizieren (siehe Koch, Gabler & Braun 1994, Schnell 1997, von der Heyde 1999). Der für die vorliegende Fragestellung wichtige Unterschied dieser Verfahren liegt in den unterschiedlichen Aufgaben/„Spielräumen“ der Interviewer bei der Selektion der Zielpersonen. Während die Interviewer bei Registerverfahren keinen Einfluss auf die Auswahl der Zielperson haben, wählen sie beim Adress-Random Design die Zielperson aus dem vorgegebenen Haushalt aus und

3 Detailliertere Ausführungen zu den ADM-Stichprobensystem und den unterschiedlichen Verfahren finden sich in z.B. in Arbeitsgemeinschaft ADM-Stichproben und Bureau Wendt (1994), Behrens & Löffler (1999), Hoffmeyer-Zlotnik (1997)

4 Bei den hier untersuchten Studien sind dies jeweils 151 Gemeinden. Zur Begründung der Anzahl von 151 Gemeinden siehe Koch & Gabler & Braun (1994: 58ff).

5 Für detaillierte Informationen zu den Einwohnermeldeamtsstichproben siehe Koch, Gabler & Braun (1994) sowie von der Heyde (1999).

bei Random-Route wählen die Interviewer sowohl den Haushalt als auch die Zielperson aus dem Haushalt aus. Die Spielräume ergeben sich einerseits, da auch umfangreiche Begehungs- und Auswahlregeln nicht alle in der Praxis vorzufindenden Situationen abdecken können, andererseits ist nicht auszuschließen, dass Interviewer die Vorgaben nicht einhalten und statt dessen eher leicht erreichbare und/oder befragungswilligere Personen interviewen werden (siehe z.B. Alt, Bien & Krebs 1991, Schnell 1997).

Aufgrund der unterschiedlich großen Spielräume der Interviewer bei der Auswahl der Zielpersonen kann bei den Registerstichproben die höchste Datenqualität erwartet werden und bei Random-Route Verfahren die geringste . Adress-Random Stichproben sollten sich zwischen den Random-Route und den Registerstichproben befinden.

2.2 Das gewählte Kriterium für die Qualität einer Stichprobe

Die Qualität der Stichproben wird anhand eines Vergleichs der Nettostichproben mit dem Mikrozensus beurteilt. Dies ist eine häufig verwendete da einfache Methode zur Untersuchung von Nonresponse (siehe z.B. Hartmann 1990, Hartmann & Schimpl-Neimanns 1992, Häder & Häder 1997, Koch 1998). Der Mikrozensus kann als „gültiges“ Außenkriterium für die Bewertung der Stichproben herangezogen werden. Die Teilnahme ist verpflichtend, so dass mit einem Unit-Nonresponse von ca. drei Prozent (Lüttinger & Riede 1997) keine über alle untersuchten Variablen hinweg gravierenden systematischen Effekte von Nonresponse zu erwarten sind. Auch Vergleiche von Randverteilungen des Mikrozensus mit anderen statistischen Informationen (Volkszählung, Hochschulstatistik) weisen gute Übereinstimmungen auf (Hartmann 1990).

Ein Nachteil der verwendeten Methode ist, dass in dieser nur für soziodemographische Variablen Aussagen getroffen werden können und nicht für die in vielen Umfragen eigentlich interessanteren inhaltlichen Variablen, da im Mikrozensus keine inhaltlichen Variablen enthalten sind und Schlussfolgerungen von den Verteilungen der soziodemographischen Variablen auf die inhaltlichen Variablen nicht oder nur begrenzt möglich sind (Schnell 1993, 1997).

In die folgende Untersuchung gehen die Merkmale Geschlecht, Alter, Bildung, Familienstand, Haushaltsgröße und Stellung im Erwerbsleben ein.

Die Stichproben haben wie zuvor beschrieben etwas andere Grundgesamtheiten, deshalb bezieht sich der Vergleich der Nettostichproben mit dem Mikrozensus nur auf die deutsche erwachsene Bevölkerung in Privathaushalten. Die untersuchten Variablen wurden auf eine vergleichbare Weise kodiert[6]. Da sich die so-

6 Zu der Kodierung der einzelnen Variablen siehe Koch (1998, S. 87ff). Für den deutschen Teil des European Social Survey wurden im Fall der Bildung die gleiche Kodierung vorgenommen wie bei den ALLBUS-Erhebungen. Im Falle der Stellung im Erwerbsleben wurden

ziodemographische Grundstruktur wandelt, wurden unterschiedliche Mikrozensen als Referenzgröße herangezogen (siehe Tabelle 1)[7]. Alle Verteilungen der Mikrozensen wurden mit den entsprechenden ZUMA zur Verfügung stehenden „Scientific Use Files" berechnet. Alle Daten aus Stichproben nach dem ADM-Design wurden auf Personenebene gewichtet berechnet.

Als Maßzahl für die Datenqualität im Fall von Verteilungen wird der Dissimilaritätsindex verwendet. Dieser ist einer in der Ungleichheitsforschung vielfach verwendeter Index, der die ungleiche Verteilung zweier Gruppen (hier Stichproben) auf relativ viele Kategorien erfasst. Der Index besticht durch einfache Interpretation: Die Maßzahl D gibt an, wie viel Prozent einer Stichprobe die Kategorien wechseln müssten, damit es zu einer gleichen Verteilung der Kategorien in beiden Stichproben kommt (Duncan & Duncan 1955). D ist definiert als:

$$D = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{A_i}{A} - \frac{M_i}{M}\right|,$$

wobei n der Anzahl der Kategorien, A und M der Anzahl der Personen der jeweiligen Stichproben und A_i bzw. M_i der Anzahl der Personen der Stichproben A bzw. B in der jeweiligen Kategorie i entspricht.

2.3 Datenqualität und Stichprobenverfahren

Zur Beurteilung der Qualität der Stichproben nach den jeweiligen Verfahren wurden für alle sechs soziodemographischen Variablen aller 26 Stichproben der jeweilige Dissimilaritätsindex berechnet.[8] In Abbildung 1 sind die durchschnittlichen Dissimilaritätsindices nach Stichprobenverfahren abgetragen. Es ist zu erkennen, dass in der Tat die Abweichungen bei den Random-Route Designs mit einem durchschnittlichen Dissimilaritätsindex von 5,7 ca. 1,8 mal höher sind als bei den Registerstichproben mit einem mittleren Wert von 3,2. Die Abweichungen bei den Adress-Random Stichproben liegen mit einem Wert von 4,7 wie erwartet zwischen den Werten der anderen Verfahren.

Personen, die im Rahmen der deutschen ESS-Erhebungen befragt wurden, als hauptberuflich erwerbstätig kodiert, wenn sie angaben, dass sie in den letzten 7 Tagen einer bezahlten Arbeit nachgingen – auch bei vorübergehender Abwesenheit, z.B. wg. Krankheit – mit einer Wochenarbeitszeit von mehr als 17 Stunden.

7 Für die Registerstichproben wurde der Mikrozensus des Vorjahres gewählt, da die Stichproben in der Regel im Jahr vor der Feldarbeit gezogen wurden.

8 Die Nettostichproben der Umfragen wurden getrennt für West- und Ostdeutschland analysiert, deshalb ergeben sich insgeamt 26 untersuchte Stichproben.

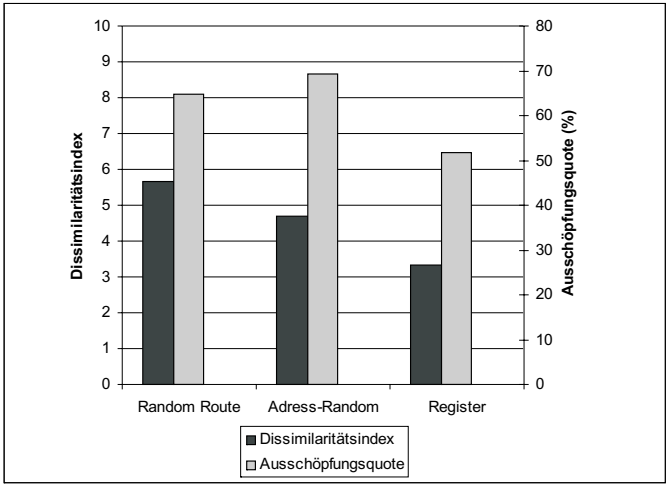


Abb. 1: Durchschnittlicher Wert der Dissimilaritätsindices und durchschnittliche Ausschöpfungsquoten nach Stichprobenverfahren

Die Unterschiede in den Stichprobenverfahren führe ich auf die unterschiedlich großen Spielräume der Interviewer bei der Auswahl der Zielpersonen zurück. Auch wenn diese These leider nicht für alle Umfragen weitergehend analysiert werden kann, so können die höheren Anteile der Verweigerungen an den systematischen Ausfällen aller ALLBUS Registerstichproben (>70%) im Vergleich zu den ALLBUS Erhebungen nach dem ADM-Design (50%-55%) sowie die in den ADM-Stichproben höheren berichteten Anteile beim ersten Kontakt realisierter Interviews (z.B. in Alt, Bien & Krebs 1991) als Indizien für das Ausweichen auf befragungsbereitere Zielpersonen gelten.

Betrachtet man sich die Abweichungen der einzelnen demographischen Informationen genauer, fällt auf, dass die Abweichungen zwischen den Registerstichproben und den ADM-Stichproben in der Regel bei den Merkmalen Haushaltsgröße und Familienstand am größten sind. Dies könnte eine Folge der Unschärfen der Transformationsgewichtung der ADM-Stichproben, z.B. durch eine nicht korrekte Erfassung der Haushaltsgröße (Rothe 1990), sein. Auch dies kann leider an dieser Stelle nicht für alle ADM-Stichproben untersucht werden. Für die untersuchten ALLBUS-Studien nach dem ADM-Design (ALLBUS 1992, 1998) verbesserte sich die Datenqualität zwar, wenn diese ungewichtet analysiert wurden, die Abweichungen blieben aber größer als die der untersuchten Registerstichproben.

2.4 Datenqualität und Ausschöpfungsquote

Neben den Abweichungen vom Mikrozensus sind in Abbildung 1 auch die durchschnittlichen Ausschöpfungsquoten nach dem Stichprobenverfahren abgetragen. Demnach weisen die Registerstichproben im Schnitt die niedrigsten Ausschöpfungsquoten auf, während die Ausschöpfungen bei den ADM-Verfahren höher ausfallen. Verschiedene Erklärungen für diesen negativen Zusammenhang ($r = -0{,}50$, $p < 0{,}01$, $N = 26$) von Datenqualität und Ausschöpfungsquote sind denkbar. Sehr plausibel ist aber eine Erklärung, welcher ein ähnlicher Mechanismus zugrunde liegt wie für den Zusammenhang von Datenqualität und Stichprobenverfahren auch. Nur in Registerstichproben ist eine (sollte eine) valide Angabe der Ausschöpfungsquoten möglich (sein).[9] Bei den anderen untersuchten Verfahren basiert die Berechnung auch immer auf den Angaben der Interviewer zu den Zielpersonen oder Zielhaushalten der Bruttostichprobe. Hier können geringe Ungenauigkeiten bei der Dokumentation der Bruttostichprobe, z.B. durch das Nichtdokumentieren einiger weniger nicht befragungsbereiter Zielpersonen, die *berichtete* Ausschöpfungsquote deutlich erhöhen (vgl. Koch 1998).

Deshalb soll in einem zweiten Schritt untersucht werden, ob der häufig postulierte positive Zusammenhang von Datenqualität und Ausschöpfungsquote zumindest bei den Registerstichproben – den Verfahren in welchen die Ausschöpfung „relativ" valide berechnet werden kann – zu finden ist. In Abbildung 2 sind die jeweiligen mittleren Dissimilaritätsindices und die Ausschöpfungsquoten der Registerstichproben abgetragen. Es ist zu erkennen, dass auch für diese Registerstichproben kein positiver Zusammenhang gegeben ist ($r = -0{,}04$, $p > 0{,}8$, $N = 14$).

Nachdem in den beiden obigen Ergebnissen ein negativer Zusammenhang bzw. kein Zusammenhang zwischen der Datenqualität und der „berichteten" Ausschöpfungsquote ermittelt werden konnte, soll weiter untersucht werden, ob zumindest innerhalb der einzelnen Umfragen ein positiver Zusammenhang zwischen Datenqualität und erreichter Ausschöpfung zu finden ist. Hierzu werden jeweils zu verschiedenen Zwischenständen der Feldarbeit, d.h. nach einer erzielten

9 Auch hier sind prinzipiell Dokumentationslücken in den meisten Einwohnermeldeamtsstichproben auszumachen, die zu einer höheren berichteten Ausschöpfung führen, so zum Beispiel die Verkodung einer Verweigerung als qualitätsneutralen Ausfall, die nicht in die Berechung der Ausschöpfung eingeht, oder aber die Subsitution von Adressen. Beides kann nur dann überprüft und ausgeschlossen werden, wenn der Auftraggeber nicht nur über die „Bruttostichprobe", d.h. über die Grundmerkmale der Zielpersonen (siehe z.B. Neller 2005), sondern auch über die Adressen der Zielpersonen verfügt und ggf. selbst Kontrollen durchführen kann – so wie dies beim ALLBUS ab 2004 der Fall ist.

Ausschöpfung von 10%, 12,5%, 15% ... usw. jeweils die Abweichungen vom Mikrozensus berechnet.[10]

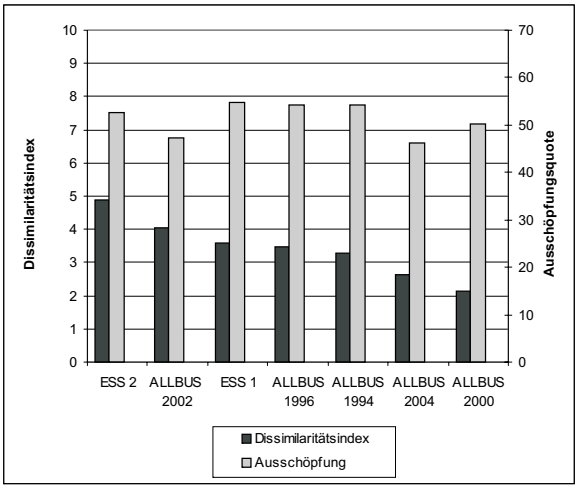


Abb. 2: Durchschnittliche Dissimilaritätsindices und Ausschöpfungsquoten der Registerstichproben

Ziel dieses Abschnitts ist es, einen ersten Einblick in die Entwicklung der Datenqualität über die Feldzeit zu bekommen. Die Hauptfragen sind, ob sich im Feldverlauf die Datenqualität verbessert und ob typische Muster einzelner Variablen festzustellen sind. Neben den univariaten Verteilungen werden im Folgenden auch bivariate und multivariate Zusammenhänge betrachtet.[11]

Dieser Teil der Untersuchung beschränkt sich auf die Registerstichproben des ALLBUS und der deutschen Teilstichproben des European Social Survey jeweils ab dem Jahr 2000.[12] Alle Analysen beziehen sich auf Gesamtdeutschland. Im Vergleich zu den Analysen zuvor wird als Grundgesamtheit die Wohnbevölkerung ab

10 Die erzielte Ausschöpfung wurde operationalisiert durch die Anzahl der realisierten Interviews zu einem Zeitpunkt t während der Feldarbeit, gemessen an der – nach der Erhebung – bereinigten Bruttostichprobe.

11 Bei der Beurteilung bivariater und multivariater Zusammenhänge werden die Differenzen zwischen den Korrelationskoeffizienten bzw. Regressionskoeffizienten, die in den Umfragen berechnet wurden, und den „Soll"-Koeffizienten aus den Mikrozensus-Erhebungen herangezogen.

12 Die Beschränkung auf diese fünf Erhebungen erfolgt, um den Erhebungszeitraum der Studien so eng wie möglich zuhalten, damit der Effekt sinkender Ausschöpfungen über die Zeit minimiert wird.

18 Jahre gewählt. Deshalb wird im Folgenden auch das Merkmal Staatsangehörigkeit (deutsch/nicht deutsch) berücksichtigt.

Problematisch an diesem Ansatz ist die Tatsache, dass verschiedene Effekte der Feldarbeit unkontrolliert in die Analysen eingehen. Um mögliche regionale Unterschiede, die durch den zeitlich versetzten Beginn der Interviewertätigkeit entstehen, zu minimieren, wurde der Tag, an dem die einzelnen Interviewer mit ihrer Arbeit begonnen hatten, auf den Feldtag Null gesetzt. Auch wurden in vier der Studien während der Feldzeit zusätzliche Adressen eingesetzt.[13] Um diesen Effekt zu minimieren, wurde der Ausgabetermin dieser Adressen auf den Feldtag Null gesetzt.

In Abbildung 3 ist die Entwicklung der mittleren Abweichungen der soziodemographischen Variablen der jeweiligen Umfragen über die Feldzeit zu erkennen. Besonders 3 Punkte erscheinen wichtig. A) Eine Verbesserung der Datenqualität über die Feldzeit ist tendenziell zu erkennen (Ausnahme ALLBUS 2002), wobei die Verbesserungen, zumindest beim ALLBUS 2000 und 2004 nach einer erzielten Ausschöpfung von ca. 30% nur noch marginal sind. Das Muster geringer werdender Veränderungen während der Feldzeit war zu erwarten, insbesondere vor dem Hintergrund der kumulierten Betrachtung der Fälle. Je später ein Fall realisiert wird, desto geringer ist dessen Einfluss – ceteris paribus – auf die Verteilungen und Zusammenhänge in der bis zu diesem Zeitpunkt realisierten Nettostichprobe. B) Die Abweichungen aller 5 Studien sind über die gesamte Feldzeit hinweg niedriger als die mittleren Abweichungen der zuvor untersuchten Stichproben nach dem ADM-Verfahren (siehe Abbildung 1). C) Es lässt sich ein Unterschied zwischen den Umfragen erkennen. Die ALLBUS Erhebungen der Jahre 2000 und 2004 weisen einen – zumindest ab einer erzielten Ausschöpfung von 25% – ähnlichen Verlauf der Datenqualität auf. Gleiches gilt für die deutschen Teilstichproben des ESS. Diese haben eine nahezu identische Entwicklung der Datenqualität. Die Abweichungen sind hier aber fast doppelt so hoch wie beim ALLBUS 2000 und 2004. Der Verlauf des ALLBUS 2002 folgt hingegen einem anderen Muster. Hier werden die Abweichungen zum Mikrozensus während der Feldphase größer und pendeln sich eher bei denen der ESS Erhebungen ein. Dies ist möglicherweise ein Hinweis darauf, dass die Abweichungen nicht nur Unterschiede der Umfrageprogramme widerspiegeln, sondern auch Folge eines Institutseffektes sein könnten. Die unterschiedlichen durchführenden Institute sind durch gestrichelte bzw. durchgezogene Linien gekennzeichnet.

13 Die Adressen wurden zusätzlich eingesetzt, damit die angstrebte Fallzahl an Interviews erreicht werden konnte.

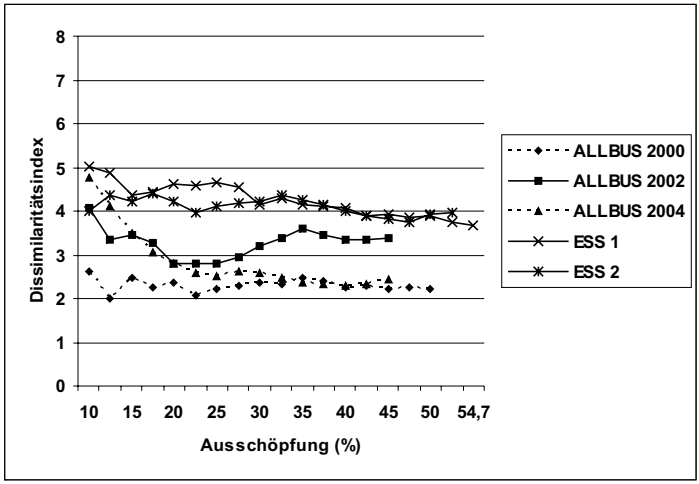


Abb. 3: Durchschnittliche Dissimilaritätsindices nach erzielter Ausschöpfungsquote während der Feldzeit, getrennt für verschiedene Umfragen

Neben der Frage der Verbesserung der Datenqualität über die Feldzeit wurden die soziodemographischen Variablen im Hinblick auf typische Muster bei den einzelnen Variablen untersucht. Drei Variablen mit typischen Verläufen über die Feldzeit konnten ermittelt werden. Dies sind die Haushaltsgröße, der Status der Erwerbstätigkeit und die Entwicklung des Merkmals Bildung. Während es bei der Haushaltsgröße und dem Erwerbsstatus über die Feldzeit in der Regel zu einer Verbesserung der Datenqualität kommt (siehe Abbildungen 4, 5) – dies ist vor dem Hintergrund der schwereren Erreichbarkeit berufstätiger Personen auch zu erwarten gewesen – nehmen die Abweichungen bei dem Merkmal Bildung in der Regel über die Feldzeit zu (siehe Abbildung 6). D.h. der u.a. von Esser, et al. (1989) und Hartmann & Schimpl-Neimanns (1992) ausgemachte Bildungsbias deutscher sozialwissenschaftlicher Umfragen nimmt über die Feldzeit zu. Dies ist auch vor dem Hintergrund der schwereren Erreichbarkeit aber zugleich höheren Kooperationsbereitschaft höherer Bildungsgruppen plausibel. Dabei sind wieder systematisch unterschiedliche Verläufe und Abweichungen zwischen den Umfragen zu erkennen. Da die Frage nach dem Schulabschluss in den ALLBUS-Befragungen immer identisch ist, scheinen die Unterschiede weniger zwischen den Umfrageprogrammen als zwischen den Instituten zu liegen.[14]

14 Zudem ist die Frage nach dem allgemeinbildenden Schulabschluss im ESS und im ALLBUS nahe zu identisch. Während im ESS gefragt wird „Was ist der höchste allgemeinbildende Schulabschluss, den sie haben?“ wird im ALLBUS gefragt „Welchen allgemeinbildenden Schulabschluss haben Sie?“ mit der Anweisung an die Interviewer „Bitte nur den höchsten

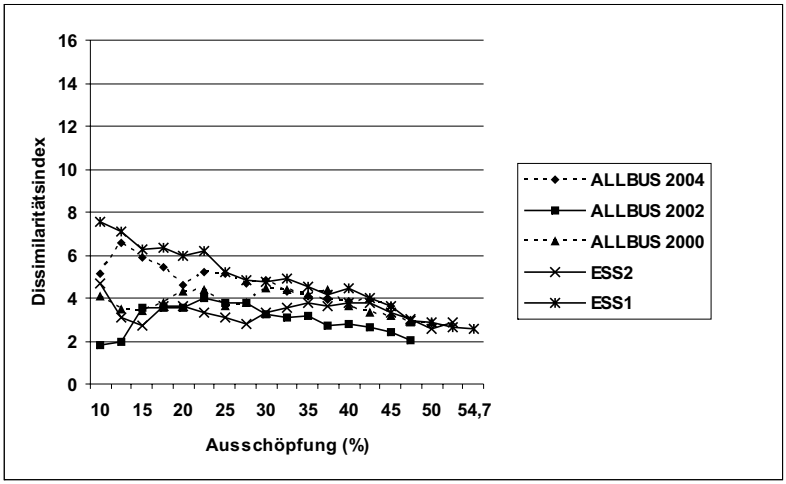


Abb. 4: Dissimilaritätsindex für das Merkmal Haushaltsgröße, nach erzielter Ausschöpfungsquote während der Feldzeit, getrennt für verschiedene Umfragen

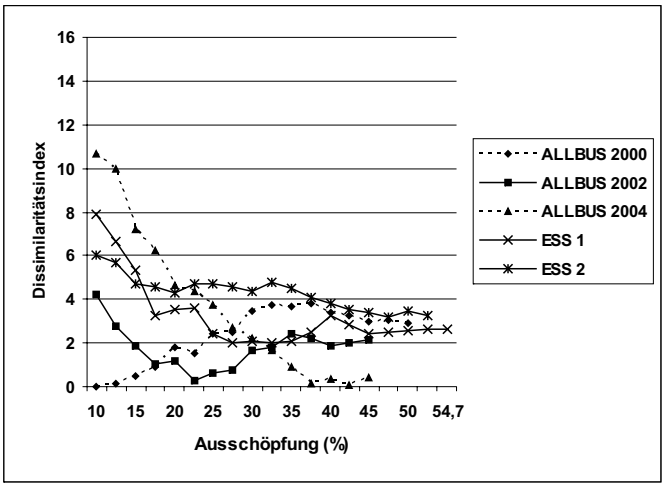


Abb. 5: Dissimilaritätsindex für das Merkmal Stellung im Erwerbsleben, nach erzielter Ausschöpfungsquote während der Feldzeit, getrennt für verschiedene Umfragen

Schulabschluss angeben lassen“. Die Antwortkategorien sind in beiden Umfrageprogrammen die gleichen.

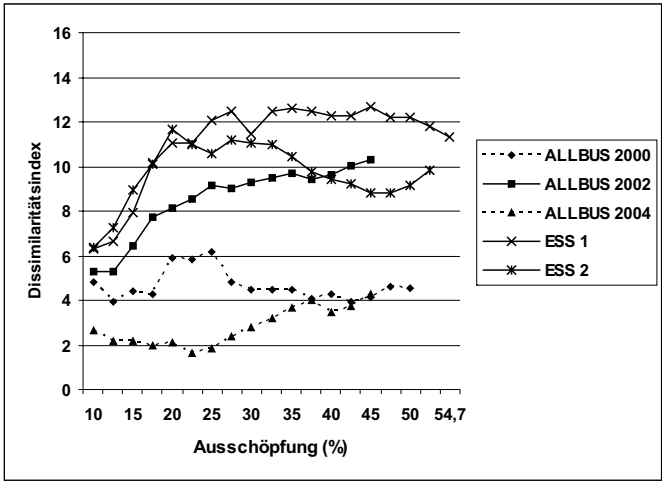


Abb. 6: Dissimilaritätsindex für das Merkmal Bildung, nach erzielter Ausschöpfungsquote während der Feldzeit, getrennt für verschiedene Umfragen

Zur Untersuchung der bivariaten Zusammenhänge wurden jeweils pro Umfrage und Zwischenstand der Feldarbeit 15 Korrelationen berechnet.[15] Die ermittelten Korrelationskoeffizienten wurden mit den „Soll-Koeffizienten" des Mikrozensus verglichen. In Abbildung 7 sind die mittleren Beträge der Abweichungen zwischen den Korrelationskoeffizienten der Umfragen und des Mikrozensus nach erreichter Ausschöpfung abgetragen. Entgegen der Erwartung verringern sich die Abweichungen zwischen den Korrelationskoeffizienten der Umfragen und des Mikrozensus bei allen Umfragen – relativ – gleichmäßig über die gesamte Feldzeit, obwohl in der kumulierten Betrachtung das relative Gewicht der später realisierten Fälle für die Korrelationen eigentlich immer geringer wird, d.h. je später die Fälle erzielt werden, umso stärker unterscheiden sie sich von den zuvor realisierten Fällen und wirken sich positiv auf die durchschnittlichen Abweichungen zum Mikrozensus aus.

Interessant ist, dass es auch im bivariaten Fall Unterschiede zwischen den Umfragen gibt. Dies könnte, wie vielleicht zunächst zu vermuten, auf das Merkmal Bildung zurückzuführen sein. Werden aber die Korrelationen mit dem Merkmal

15 Dies waren im Einzelnen die bivariaten Korrelationen der Merkmale Geschlecht x (Alter, Bildung, Familienstand, Haushaltsgröße, Stellung im Erwerbsleben), Alter x (Bildung, Familienstand, Haushaltsgröße, Stellung im Erwerbsleben), Bildung x (Familienstand, Haushaltsgröße, Stellung im Erwerbsleben), Stellung im Erwerbsleben x (Familiestand, Haushaltsgröße) und Haushaltsgröße x Familienstand.

Bildung nicht berücksichtigt, bleiben die Unterschiede zwischen den Umfragen bestehen.

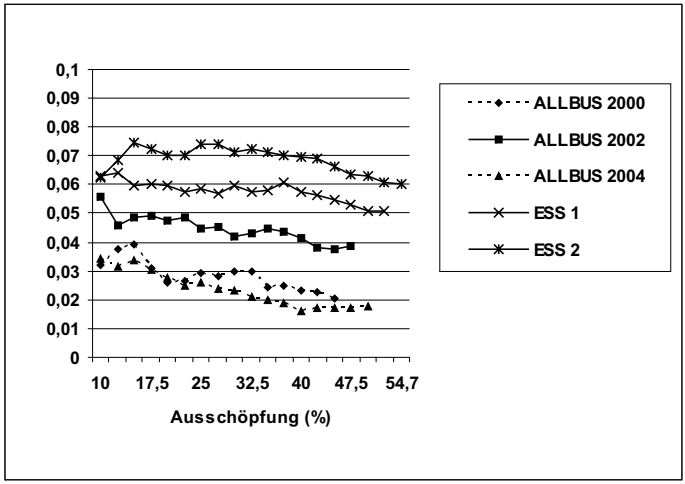


Abb. 7: Durchschnittliche Beträge der Abweichungen der Korrelationskoeffizienten der Umfragen und des Mikrozensus, nach erzielter Ausschöpfung über die Feldzeit

Zur Untersuchung der multivariaten Zusammenhänge wurden logistische Regressionsmodelle berechnet. Für jede Umfrage wurde zu den unterschiedlichen Ausschöpfungszwischenständen jeweils das Merkmal Erwerbstätigkeit durch die kategorialen Variablen Geschlecht, Alter und Bildung erklärt.[16] Zur Beurteilung der Modelle wurde der mittlere Betrag der Differenzen der Konstanten und der Regressionskoeffizienten zu dem jeweiligen „Soll-Modell" des Mikrozensus ermittelt.[17] Die berechneten mittleren Beträge der Abweichungen nach Studie und Feldzwischenstand finden sich in Abbildung 8. Auch in diesem multivariaten Fall überraschen die Ergebnisse. Zum einen finden sich hier über die gesamte Feldzeit mehr oder weniger starke Veränderungen der Datenqualität. Zum anderen ist erstaunlich, dass auch in diesem Fall systematische Unterschiede zwischen den

16 Die Analysen sind auf die unter 60-Jährigen beschränkt.

17 Zum Vergleich der Modelle hätten auch die in den Regressionsmodellen geschätzten Anteilswerte Erwerbstätiger analysiert werden können. Dies wäre aber aufgrund der zahlreich zu schätzenden Anteilswerte pro Umfrage zu unübersichtlich, deshalb wurden näherungsweise die Abweichung der Koeffizienten und der Konstanten berücksichtigt.

Umfrageinstituten und weniger zwischen den Umfragen zu finden sind.[18] Die ermittelten Beträge der Abweichungen der Umfragen pro Institut gleichen sich über die Feldzeit an, aber auf einem unterschiedlichem Niveau.

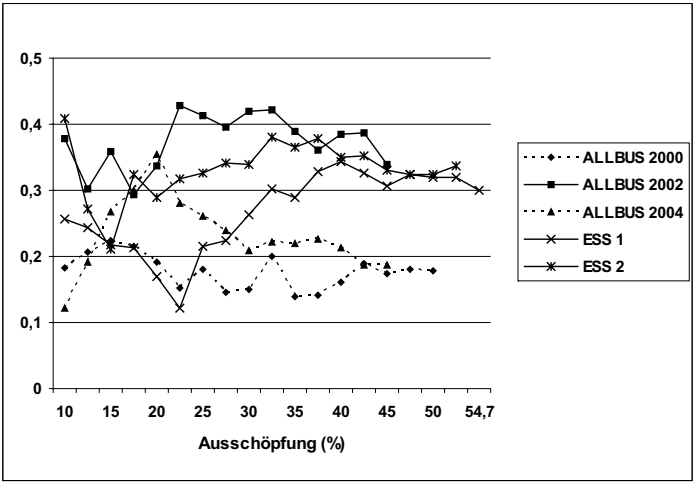


Abb. 8: Durchschnittliche Beträge der Abweichungen der Regressionskoeffizienten und Konstanten zwischen den Umfragen und des Mikrozensus

3 Fazit

Es konnte gezeigt werden, dass sich die Datenqualität von nationalen allgemeinen Bevölkerungsumfragen je nach eingesetzten Stichprobenverfahren unterscheidet. In den untersuchten Random-Route Stichproben sind die Abweichungen in der soziodemographischen Grundstruktur nahezu doppelt so hoch wie in den untersuchten Registerstichproben. Es zeigt sich ferner kein Zusammenhang zwischen der Datenqualität und den Ausschöpfungsquoten der untersuchten Umfragen. Wird die Datenqualität der Registerstichproben im Verlauf der Feldarbeit analysiert, zeigt sich für die demographische Grundsstruktur hinsichtlich der un-

18 Im ALLBUS werden die untersuchten Variabeln immer auf die gleiche Art erfragt. Da der ALLBUS 2002 sich im Bereich der ESS-Studien bewegt, scheint auch dies Ergebnis eher auf einen Institutseffekt hinzuweisen. Bei näherer Betrachtung der Koeffizienten ist zu erkennen, dass in den ESS-Erhebungen und im ALLBUS 2002 insbesondere die Erwerbstätigkeit der Altergruppen der 18- bis 30-Jährigen im Vergleich zur Erwerbstätigkeit der 50- bis 60-Jährigen unterschätzt wird.

tersuchten Verteilungen, der Korrelationen und des multivariaten Zusammenhangs, dass sich die Abweichungen zum Mikrozensus über die Feldzeit im Allgemeinen reduzieren. Es konnten bei den Registerstichproben Unterschiede zwischen den Umfragen ausgemacht werden, die eher auf eine unterschiedliche Durchführung des jeweiligen Umfrageinstitutes zurückzuführen sind als auf Unterschiede zwischen den Umfrageprogrammen ALLBUS und ESS an sich.

Alles in allem zeigen die hier vorgestellten Ergebnisse, dass es einen sicheren Weg zu qualitativ hochwertigen Umfragedaten vielleicht in der Theorie, aber sicher nicht in der Praxis gibt. Eine ADM-Stichprobe nach Random-Route-Design sollte rein theoretisch genauso gut oder schlecht die Grundgesamtheit abbilden wie eine Registerstichprobe. In der Praxis scheinen die Registerstichproben doch überlegen zu sein. Die Wahrscheinlichkeit von Verzerrungen der realisierten Stichprobe sinkt eigentlich mit steigender Ausschöpfungsquote. Weil aber ein Ausfall eben kein zufälliges Ereignis ist, kann letztendlich doch eine Studie mit 50% Ausschöpfung einer ganz ähnlich angelegten Studie mit 60% Ausschöpfung nach den verfügbaren Außenkriterien überlegen sein. Allerdings deuten die Ergebnisse insgesamt, insbesondere die zu den bivariaten Zusammenhängen, doch darauf hin, dass man mit zunehmender Ausschöpfung tendenziell „der Wahrheit" – zumindest was die hier betrachteten soziostrukturellen Variablen angeht – immer näher kommt. Offen bleibt jedoch, wie andere inhaltliche Variablen, v.a. auch Einstellungsvariablen, mit der Ausschöpfungsquote bzw. einzelnen Aspekten des Ausfallgeschehen – Teilnahmebereitschaft, Erreichbarkeit, Befragungsfähigkeit – zusammenhängen und inwieweit dementsprechend bei diesen Variablen das Bemühen um eine hohe Ausschöpfung „lohnt".

Literatur

Alt, Christian, Bien, Walter und Krebs, Dagmar, (1991). Wie zuverlässig ist die Verwirklichung von Stichprobenverfahren? Random Route versus Einwohnermeldeamtsstichprobe. ZUMA-Nachrichten, 28: 65-72.

Arbeitsgemeinschaft ADM-Stichproben und Bureau Wendt (1994). Das ADM-Stichproben-System. Stand 1993. In: S. Gabler/J.H.P. Hoffmeyer-Zlotnik/D. Krebs (Hrsg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag. S. 188-202.

Behrens, Kurt, Löffler, Ute (1999). Aufbau des ADM-Stichproben-Systems. In: ADM Arbeitsgemeinschaft Deutscher Markt- und Sozialforschungsinstitute e.V. und AG.MA Arbeitsgemeinschaft Media-Analyse e.V. (Hrsg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis. Opladen: Leske + Budrich. S. 69-91.

Curtin, Richard, Presser, Stanley, Singer, Eleanor (2000). The Effects of response Rate Changes on the Index of Consumer Sentiment. In: Public Opinion Quarterly, Vol. 64. S. 413-428.

Davis, James A., Mohler, Peter Ph., Smith, Tom W. (1994). Nationwide General Social Surveys. In: Borg, Ingwer, Mohler, Peter Ph. (Hg.): Trends and Perspectives in Empirical Social Research. Berlin: de Gryter, S. 17-25.

Duncan, Otis Dudley, Duncan Beverly (1955). A Methodological Analysis of Segregation Indexes. American Sociological Review 20:210 - 217.

Esser, Hartmut, Heinz Grohmann, Walter Müller, Karl August Schäffer (1989). Mikrozensus im Wandel. Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung, Band 11 der Schriftenreihe Forum der Bundesstatistik, Stuttgart.

Häder, Michael, Häder, Sabine (1997). Adreßvorlaufverfahren: Möglichkeiten und Grenzen. Eine Untersuchung am Beispiel der Erhebung Leben Ostdeutschland1996. S. 43- 67, in: Gabler, Siegfried und Hoffmeyer-Zlotnik, Jürgen H. P. (Hrsg.), Stichproben in der Umfragepraxis. Westdeutscher Verlag GmbH, Opladen.

Hartmann, Peter, (1990). Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich das ALLBUS und des Mikrozensus. ZUMA-Nachrichten, 26, S. 7-30.

Hartmann, Peter; Schimpl-Neimanns, Bernhard (1992). Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44, S. 315-346.

Hoffmeyer-Zlotnik, Jürgen H. P., (1997). Random-Route-Stichproben nach ADM. S. 33- 42, in: Gabler, Siegfried und Hoffmeyer-Zlotnik, Jürgen H. P. (Hrsg.), Stichproben in der Umfragepraxis. Westdeutscher Verlag GmbH, Opladen.

Keeter, Scott, Miller, Carolyn, Kohut, Andrew, Groves, Robert M., Presser, Stanley (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. In: Public Opinion Quarterly, Vol. 64. S. 125-148.

Koch, Achim, (1998). Wenn „Mehr“ nicht gleichbedeutend mit „Besser“ ist: Ausschöpfungsquoten und Stichprobenverzerrungen in Allgemeinen Bevölkerungsumfragen. ZUMA-Nachrichten, 42: 66-90.

Koch, Achim, Gabler, Siegfried, Braun, Michael (1994). Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften” (ALLBUS) 1994. ZUMA-Arbeitsbericht 94/11

Lüttinger, Paul, Riede, Thomas (1997). Der Mikrozensus. Amtliche Daten für die Sozialforschung. ZUMA-Nachrichten 41: S.19-43.

Merkle, Daniel, Edelman, Murray. (2002). „Nonresponse in Exit Polls: A Comprehensive Analysis.“ In Survey Nonresponse, (Hrsg.) R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, S. 243–58. New York: Wiley.

Neller, Katja (2005). Kooperation und Verweigerung: Eine Non-Response-Studie, in ZUMA-Nachrichten 57, S. 9-36.

Rothe, Günter, (1990). Wie (un)wichtig sind Gewichtungen? Eine Untersuchung am ALLBUS 1986. ZUMA-Nachrichten, 26: S. 31- 55.

Schnell, Rainer, (1997). Nonresponse in Bevölkerungsumfragen – Ausmaß, Entwicklung und Ursachen. Leske+Budrich, Opladen.

Schnell, Rainer, (1993). Homogenität sozialer Kategorien als Vorraussetzung für „Repräsentativität“ und gewichtungsverfahren. Zeitschrift für Soziologie, 1, 22 S. 16-32.

von der Heyde, Christian, (1999). Sonderstichproben. S. 113-123, in: ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (Hrsg.), Stichproben-Verfahren in der Umfrageforschung. Opladen: Leske+Budrich.

Registergestützter Zensus – Aktueller Stand und Entwicklungsperspektiven

Sonja Krügener

Abstract

2001 wurden die wesentlichen Verfahrensbestandteile getestet, die 2011 für einen registergestützten Zensus in Deutschland angewendet werden sollen. Ziel dieses Tests war es, Erfahrungen mit der Wirkungsweise der Module zu sammeln und Daten über die Qualität der Melderegister der Gemeinden zu erhalten. Es zeigte sich, dass die Fehler der Melderegister nicht zufällig streuen, sondern in Abhängigkeit von der Größe der Gemeinde. Die Fehlerraten konnten durch Anwendung der verschiedenen Module zwar gesenkt werden, dennoch blieb die Abhängigkeit von der Gemeindegröße. Daher wird das Verfahren um eine Stichprobe erweitert, die zum einen der Abschätzung der Fehler der Melderegister dient und zum anderen der Erhebung weiterer Merkmale, die nicht in Registern enthalten sind. Im Gegensatz zu einem traditionellen Zensus wird der registergestützte deutlich kostengünstiger und weniger belastend für die Bevölkerung sein. Unter Berücksichtigung von statistischen Fehlern wird es kleinräumige Ergebnisse für demographische Daten, Haushaltszusammenhänge sowie teilweise erwerbsstatistische Daten geben. Kleinräumige Auswertungen werden hingegen nicht möglich sein für Daten, die nicht aus Registern oder der Gebäude- und Wohnungszählung, sondern aus der Stichprobe zu beziehen sind.

1 Einführung

2011 wird Deutschland voraussichtlich an der EU-weiten Zensusrunde teilnehmen. Im Gegensatz zu den bisherigen traditionellen Zensen wird dieser wahrscheinlich registergestützt ablaufen, also die kommunalen Melderegister als Hauptquelle haben. Erste Überlegungen zu einem registergestützten Zensus gab es bereits Anfang der 90er Jahre. Da die Melderegister jedoch als zu fehlerhaft angesehen wurden und die technischen Möglichkeiten noch nicht ausgereift genug waren, wurde die Idee damals noch verworfen. Auch für die EU-weite Zensusrunde 2001, an der sich die amtliche Statistik in Deutschland eigentlich beteiligen wollte, war ein traditioneller Zensus mit Stichprobe zur Erfassung weiterer Merkmale vorgesehen (das gleiche Verfahren wurde bei der Volkszählung 1970

angewendet). Dieses Verfahren fand jedoch wegen der Widerstände in der Bevölkerung im Vorfeld der Volkszählung 1987 in der Politik keine Unterstützung, so dass es 1996 zur politischen Entscheidung gegen eine traditionelle Zählung kam. Die Daten sind jedoch unverzichtbar, da Zensen die Grundlage für die Bevölkerungszahlen der amtlichen Statistik und damit für vielfältige Planungen in Politik, Wirtschaft und Wissenschaft sind. Es sollte daher ein Methodenwechsel hin zu einem registergestützten Zensus eingeleitet werden. Für einen derartigen Zensus fehlten in Deutschland jedoch jegliche Erfahrungen, so dass weit reichende Untersuchungen notwendig waren und Deutschland schon aus zeitlichen Gründen nicht an der Zensusrunde 2001 teilnehmen konnte. Es wurde stattdessen zum Stichtag 5. Dezember 2001 ein groß angelegter Register- und Verfahrenstest durchgeführt, um die Machbarkeit eines registergestützten Zensus zu überprüfen.

2 Der Zensustest und seine Ergebnisse

Der Test diente zwei Zielen (siehe Abb. 1): der Überprüfung der Qualität der Melderegister (Registertest) sowie der Durchführbarkeit von verschiedenen Verfahrensmodulen (Verfahrenstest). Um die Qualität der Melderegister zu überprüfen, wurden bundesweit in verschiedenen Gemeinden durch eine geschichtete Stichprobe Melderegisterauszüge gezogen (rund 38.000 Adressen). Zur Überprüfung der Richtigkeit der Angaben in diesen Adressen wurde eine Haushaltebefragung durchgeführt. Auf diese Weise konnten Zahlen über Karteileichen (Übererfassungen) und Fehlbestände (Untererfassungen) ermittelt werden. Ein zweiter Melderegisterauszug vier Monate später diente dazu, temporäre Registerfehler herauszufiltern. Temporäre Registerfehler kommen dann zustande, wenn sich Zugezogene zum Stichtag noch nicht an- oder Weggezogene noch nicht abgemeldet haben.

Für den Verfahrenstest wurden in einer Unterstichprobe (16.000 Adressen) bei der Haushaltebefragung zusätzliche Fragen gestellt, um so die Anwendung verschiedener weiterer Verfahrensmodule zu testen. Zu den zu testenden Verfahrensmodulen gehörten die postalische Gebäude- und Wohnungszählung, die Haushaltegenerierung und eine Zusammenführung mit den Daten der Bundesanstalt, heute Bundesagentur für Arbeit. In der Gebäude- und Wohnungszählung wurden die Eigentümer nach Angaben zu Gebäuden und Wohnungen, u. a. zum Wohnungsinhaber (dem Selbstnutzer oder Hauptmieter), befragt. Die Frage nach dem Wohnungsinhaber diente in erster Linie der Haushaltegenerierung. Da die Melderegister nur Angaben für Adressen und nicht für einzelne Wohneinheiten haben, wurden mit Hilfe der Gebäude- und Wohnungszählung Haushalte generiert. Dazu wurden die Informationen über die Anzahl der bewohnten Wohnungen je Gebäude und den Wohnungsinhaber genutzt. So wurden die mit den Wohnungs-

inhabern verknüpften Wohnungen mit den aus den Melderegistern ermittelten zusätzlichen Personen belegt. Kriterien für die zu bildenden Haushalte waren u. a. die Verzeigerungen zu Ehepartnern und Kindern in den Melderegistern, gleicher Name sowie gleiches Einzugsdatum. Zu den auf diese Weise generierten Haushalten wurden in der Zusammenführung die Daten der Bundesagentur für Arbeit zugefügt, so dass am Ende Einzeldatensätze mit demographischen, haushalts-, wohnungs- und erwerbsstatistischen Angaben vorlagen, die mit Hilfe der Haushaltebefragung überprüft werden konnten. Losgelöst von den Register- und Verfahrenstests wurde ein weiteres Verfahrensmodul, die Mehrfachfallprüfung, getestet. Diese dient dazu, Personen, die mehrfach mit Hauptwohnsitz gemeldet sind, herauszufinden. Dieser Test bedurfte eines anderen Registerauszuges als die anderen beiden Tests, da es wichtig war, Personen bundesweit ausfindig zu machen, die mehrfach gemeldet waren. Daher wurden aus allen Gemeinden der Bundesrepublik Registerauszüge von Personen gezogen, die entweder an drei bestimmten Tagen Geburtstag hatten oder mit unvollständigem Geburtsdatum gemeldet waren.

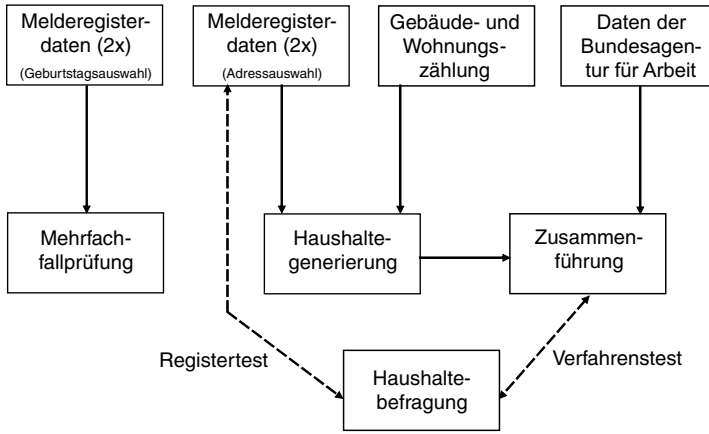


Abb. 1: Zensustest 2001

Als Ergebnis des Zensustests kann festgehalten werden, dass sich die Verfahrensmodule bewährt haben, auch wenn teilweise Weiterentwicklungsbedarf offenbar wurde. Ein Problem stellen die Karteileichen und Fehlbestände dar. Zwar konnte durch das Herausfiltern der temporären Karteileichen, durch die Mehrfachfallprüfung und durch die Haushaltegenerierung die Karteileichenrate stark verringert werden, doch auch nach diesen Verbesserungen waren diese Fehlerraten zumindest in Gemeinden mit mehr als 10.000 Einwohnern höher als der angestrebte Wert von +/- 1%. Die Fehler streuten zudem in Abhängigkeit von der Größe der

Gemeinde (siehe Tabelle 1). Die Hauptaufgabe der Weiterentwicklung des Zensusmodells besteht daher darin, die Registerfehler weiter zu senken. Dies soll mit einer Stichprobenergänzung erreicht werden.

Tab. 1: Registerfehler nach Gemeindegrößenklassen

	Gemeinden bis 10.000 Einwohner	Gemeinden mit 10 – 50.000 Einwohnern	Gemeinden mit 50 –100.000 Einwohnern	Gemeinden mit 100.000 Einwohnern und mehr	Deutschland
Karteileichenrate der Melderegister	2,8 %	3,4 %	3,8%	5,9 %	4,1 %
Fehlbestandsrate der Melderegister (ohne temp. FB)	1,3 %	1,3 %	2,1 %	2,4 %	1,7 %
Karteileichenrate (ohne temp. KL, nach MFFP und HHGen)	0,7 %	1,4 %	1,5 %	3,4 %	1,8 %

3 Aktuelles Modell eines registergestützen Zensus

Das aktuelle, auf den Zensustestergebnissen aufbauende Verfahren (siehe Abb. 2) besteht aus allen im Zensustest getesteten Modulen zuzüglich der Erweiterung um eine Stichprobe in allen Gemeinden über 10.000 Einwohnern. Die Stichprobe dient zwei Zielen: zum einen kann durch sie eine Registerfehlerabschätzung je Gemeinde vorgenommen werden und die aus den Melderegistern stammenden demographischen Daten entsprechend statistisch korrigiert werden. Die Stichprobe kann zudem genutzt werden, um weitere Merkmale zu erheben, die nicht aus Registern zu entnehmen sind, wie Pendlerverhalten, Bildung, Angaben zu Selbständigen. Dazu ist eine relativ kleine Erweiterung des Stichprobenumfangs notwendig. Diese zusätzlichen Daten werden im Gegensatz zu den Daten aus Registern und Gebäude- und Wohnungszählung nicht flächendeckend vorliegen, sondern als Schätzungen für die jeweilige Gemeinde. Voraussichtlich wird die Stichprobenerhebung so gestaltet werden, dass neben Ergebnissen für Gemeinden über 10.000 Einwohner auch Ergebnisse auf Kreis- und Gemeindeverbandsebene sowie für Stadtbezirke der größten Städte zu erhalten sind. Damit würde den Anforderungen der EU für diese Merkmale Genüge getan, die derartige Ergebnisse auf Kreisebene (NUTS 3) fordert.

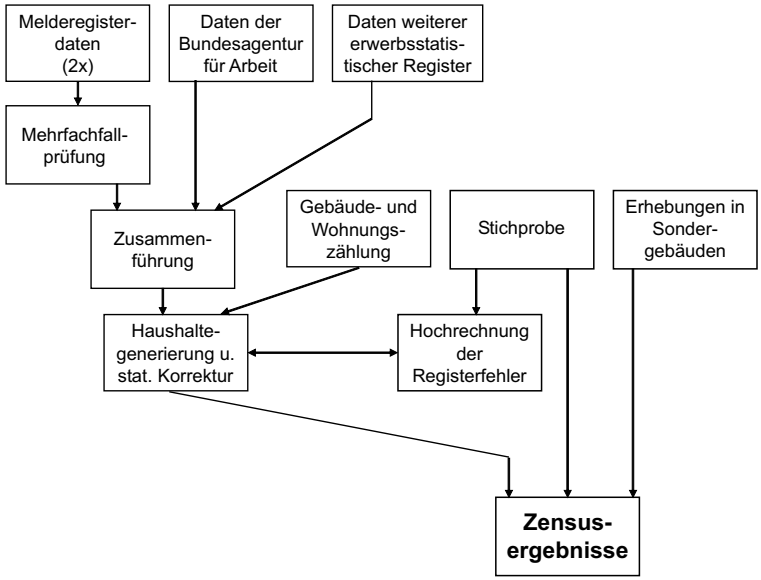


Abb. 2: Das aktuelle Zensusmodell

Dass die Stichprobe nur in Gemeinden vorgenommen wird, die mehr als 10.000 Einwohner haben hat seine Ursache in Kosten-/Nutzenüberlegungen. Die Ergebnisse der Gemeindegrößenklasse unter 10.000 Einwohnern haben nach den Bereinigungen schon eine akzeptable Qualität (siehe Tabelle 1). Durch den verringerten Stichprobenumfang sind deutlich weniger Personen von der Befragung betroffen. So wären statt 20,4 Mio. nur 7,6 Mio. Personen betroffen. In sehr kleinen Gemeinden hätte ein entsprechender Stichprobenumfang eine Vollerhebung der gesamten Gemeinde zur Folge. Eine geringere Gesamtstichprobe würde natürlich auch deutlich geringere Kosten mit sich bringen (336 Mio. statt 454 Mio.). Derzeit wird noch untersucht, inwieweit es möglich ist die Ergebnisse für Gemeinden unter 10.000 Einwohnern zu verbessern, insbesondere um die tendenzielle Untererfassung in kleinen Gemeinden (Fehlbestände 1,3%, Karteileichen 0,7%) zu beseitigen, sowie mit Hilfe von Small Area Schätzungen auch Ergebnisse bezüglich der zusätzlichen Merkmale für kleine Gemeinden aus den Kreisergebnissen zu erhalten.

Das aktuelle Modell würde statistische Einzeldatensätze liefern. Dadurch ist die Möglichkeit gegeben Einzeldatensätze frei zu aggregieren. Dabei muss berücksichtigt werden, dass der relative Zufallsfehler umso größer ist, je kleiner die auszuwertende Einheit gewählt wird. Es werden noch Kriterien entwickelt, nach denen eine Auswertung unter statistischen Gesichtspunkten durchführbar ist.

4 Traditionelle Volkszählung versus registergestützter Zensus

Wie bei einem traditionellen Zensus liegen alle Daten, die aus Registern zu beziehen sind, also im Wesentlichen die demographischen Merkmale aus den Melderegistern und die erwerbsstatischen Angaben von der Bundesagentur für Arbeit, flächendeckend vor. Durch die Stichprobe werden die Ergebnisse für Gemeinden über 10.000 Einwohner recht genau sein. In Gemeinden unter 10.000 Einwohnern sind die Register verhältnismäßig gut, jedoch muss hier mit einer nicht zu schätzenden Streuung gerechnet werden, die eventuell durch derzeit in der Entwicklung befindliche Verfahren noch verringert werden kann. Bezüglich der Registerdaten kommen die Ergebnisse recht nah an die einer klassischen Volkszählung heran, insbesondere auf Gemeindeebene. Den kleinräumigen Auswertungsmöglichkeiten sind in ihren Möglichkeiten jedoch Grenzen gesetzt.

Nachteilig an einem registergestützten Zensus ist die Tatsache, dass flächendeckend und als Einzeldaten nur Merkmale vorliegen, die auch tatsächlich bundesweit in den Melderegistern, in den Registern der Bundesagentur für Arbeit und eventuell weiteren Registern geführt werden. Nicht unbedeutende Angaben wie Religionszugehörigkeit oder Migrationshintergrund sind nicht oder nicht in der gewünschten Form aus den Melderegistern zu ziehen, genauso wenig wie Daten über Selbständige aus den Registern der Bundesagentur für Arbeit. Diese Angaben lassen sich nur aus der Stichprobe entnehmen, sodass nur Abschätzungen für Gemeinden über 10.000 Einwohner möglich sind bzw. je nach Stichprobendesign auch auf Kreis-, Gemeindeverbands- und Stadtbezirksebene.

Daten zu Gebäuden und Wohnungen werden durch die postalische Gebäude- und Wohnungszählung in jeder beliebigen regionalen Gliederung vorliegen und stehen damit den Auswertungsmöglichkeiten einer klassischen Volkszählung in nichts nach. Was die Qualität der Angaben angeht kann vermutet werden, dass die Angaben, die von den Eigentümern gemacht werden, nicht schlechter sind als diejenigen der Bewohner bei einem traditionellen Zensus.

Ein wesentlicher Vorteil des registergestützten Zensus sind die deutlich geringeren Kosten gegenüber einem traditionellen Zensus. Nach ersten Kalkulationen würde ein traditioneller Zensus in etwa eine Milliarde kosten. In die Höhe werden die Kosten insbesondere durch die Personalkosten für die Interviewer getrieben. Der registergestützte Zensus mit Stichprobe würde 2011 in etwa nur ein Drittel kosten. Darin enthalten sind bereits die Entwicklungskosten, die bei nachfolgenden registergestützten Zensen unter Umständen gesenkt werden könnten. Ein weiterer Faktor ist die geringere Belastung für die Bevölkerung. Durch die Stichprobe werden in etwa 7,6 Mio. Personen direkt befragt werden zuzüglich der postalischen Befragung der Gebäude- und Wohnungseigentümer, also ein Bruchteil der Personen, die bei einer Vollerhebung befragt würden (ca. 82 Mio.).

5 Zeitplan und Ausblick

Der Zensus 2011 wird gemeinsam von den statistischen Ämtern des Bundes und der Länder in Zusammenarbeit mit Vertretern aus der Wissenschaft vorbereitet. Das Jahr 2011 wird voraussichtlich von der EU vorgegeben, da um diesen Zeitraum allen EU-Mitgliedsstaaten ein Zensus angeordnet wird, um vergleichbare Daten zu erhalten. Diesbezüglich liegen erste Entwürfe für eine Verordnung vor, die einen Zensus alle zehn Jahre für die Mitgliedstaaten verpflichtend macht, wobei die Methode, ob nun traditionell oder registergestützt, den Mitgliedstaaten freigestellt ist. Der Zeitrahmen bis 2011 ist sehr eng bemessen. So muss die Methodenentwicklung bereits deutlich vorher abgeschlossen sein, um darauf aufbauend mit den organisatorischen Vorbereitungen beginnen zu können. Zudem muss ein Adressregister für die Durchführung der postalischen Gebäude- und Wohnungszählung eingerichtet werden und eine nationale Gesetzesgrundlage sowohl für die Erhebung als auch für den Aufbau eines Adressregisters geschaffen werden. Die ersten Ergebnisse werden voraussichtlich ein Jahr nach Ablauf des Erhebungsjahres an die EU zu liefern sein. Dieser Zeitraum wird auch benötigt werden auf Grund der umfangreichen Registerzusammenführungen und -abgleiche, die durch das komplexe Verfahren zustande kommen. Der registergestützte Zensus 2011 wird dann der erste seiner Art in Deutschland sein. Er wird den Weg bereiten für kommende Zensen in Deutschland, die auf den Erkenntnissen des Zensus 2011 aufbauen können und weiterentwickelt werden.

6 Literatur

Grohmann, H. (2000): Geschichte und Zukunft der Volkszählung in Deutschland. In: Berliner Statistik, Monatschrift 7 - 12/00, Berlin, S. 216 - 223.

Grohmann, H.; Sahner, H. und Wiegert, R. (Hrsg.) (1999): Volkszählung 2001 – Von der traditionellen Volkszählung zum Registerzensus. In: Sonderhefte zum allg. statistischen Archiv, Heft 33, Göttingen: Vandenhoeck & Ruprecht.

Schäfer, J. (2004): Ergänzende Verfahren für einen künftigen registergestützten Zensus. In: Landesamt für Datenverarbeitung und Statistik NRW (Hrsg.): Statistische Analysen und Studien, Bd. 17, Düsseldorf, S. 28 - 46.

Statistische Ämter des Bundes und der Länder (2004): Ergebnisse des Zensustests. In: Landesamt für Datenverarbeitung und Statistik (LDS) NRW (Hrsg.): Statistische Analysen und Studien, Bd. 17, Düsseldorf, S. 20 - 27.

Statistisches Bundesamt (Hrsg.) (1992): Volkszählung 2000 – oder was sonst? In: Forum der Bundesstatistik, Bd. 21, Wiesbaden.

Egbert, H.; Forster, M.; Schäfer, J.; Scharnhorst, S.; Scharmer, M. (2002): Deutschland auf dem Weg zum registergestützten Zensus. In: LDS NRW (Hrsg.): Statistische Analysen und Studien, Bd. 4, Düsseldorf.

Der Berner Stichprobenplan
Ein Vorschlag für eine effiziente Klumpenstichprobe am Beispiel der Schweiz[1]

Ben Jann

Die meisten Bevölkerungsumfragen in der Schweiz beruhen auf Stichproben, die aus dem Telefonregister gezogen werden. Da solche Stichproben bezüglich der Abdeckung der Grundgesamtheit als problematisch anzusehen sind, diskutiere ich hier ein alternatives, von Fritschi et al. (1976) entwickeltes Stichprobenverfahren, bei dem in einem ersten Schritt Gemeinden ausgewählt und dann die Adressen der Zielpersonen über die Einwohnerregister bestimmt werden. Das Verfahren ist dabei so angelegt, dass die Stichprobe ähnlich wie bei einer einfachen Klumpenstichprobe auf eine relativ geringe Anzahl Gemeinden verdichtet wird, die Stichprobe aber trotzdem eine möglichst hohe statistische Effizienz beibehält. Eine Analyse der theoretischen Eigenschaften des Berner Stichprobenplans zeigt, dass das ursprüngliche Verfahren zu leicht verzerrten Stichproben führt. Eine korrigierende Modifikation des Verfahrens wird vorgeschlagen. Zudem wird ein alternativer Ansatz mit verbesserten Eigenschaften, die *ex ante* geteilte Stichprobe, vorgestellt. Mit Hilfe einer Simulationsstudie werden sodann die Vorzüge des Berner Stichprobenplans gegenüber einer einfachen Klumpenstichprobe illustriert.

1 Einleitung

Wie auch in anderen Ländern ohne zentrales Einwohnerregister stellt in der Schweiz die Ziehung einer repräsentativen Bevölkerungsstichprobe die empirische Sozialforschung vor schwerwiegende Probleme. Ohne Urliste der Mitglieder der zu untersuchenden Population lässt sich naturgemäß nur schwer eine unverzerrte Zufallsstichprobe aus derselben ziehen. Zudem ist es, wenn man über ein ansprechendes und praktikables Stichprobenverfahren verfügt, auch nicht immer ganz einfach festzustellen, ob das Verfahren tatsächlich unverzerrte Resultate liefert. Die Erstellung von qualitativ hochwertigen Stichproben kann also sehr zeitaufwändig sein und unter Umständen erhebliche Kosten verursachen. Somit ist wohl nicht zuletzt aufgrund beschränkter Forschungsbudgets damit zu rech-

1 Im Gedenken an Herbert Iff († 18. April 1998).

nen, dass ein großer Teil der empirischen Umfrageforschung in der Schweiz auf zumindest zweifelhaften Stichproben beruht.

Die einfachste und wahrscheinlich am Häufigsten angewandte Stichprobenmethode in der Schweiz ist die Ziehung einer Haushaltsstichprobe aus dem Telefonverzeichnis (vgl. etwa Jann 2001; für Deutschland Schnell 1997), die dann in einem zweiten Schritt nach der Erfassung der Haushaltsstrukturen durch Ziehung von einzelnen Zielpersonen in den Haushalten in eine Personenstichprobe überführt wird. Die Probleme dieser Methode sind bekannt: Erstens gibt es einen gewichtigen Anteil der Bevölkerung, der auf diesem Weg nicht erreicht werden kann, weil kein oder nur ein nicht-registrierter Telefonanschluss vorliegt (zu etwas älteren Schätzungen dieses Anteils siehe zum Beispiel die Studien von Schmugge und Grau 1998, 2000; Experten gehen heute von einem Anteil nicht über das Telefonverzeichnis erreichbarer Personen von 10 bis 15 Prozent aus). In der Schweiz gibt es nach wie vor Haushalte ohne Telefonanschluss und seit 1998 kann die sonst automatische Aufnahme eines Festnetzanschlusses ins öffentliche Telefonverzeichnis verweigert werden. Weiterhin werden Mobiltelefonanschlüsse nur auf speziellen Wunsch hin in das Verzeichnis aufgenommen, was eine entsprechend geringe Registrierungsquote zur Folge hat. Zweitens handelt es sich bei der Ziehung aus dem Telefonregister wie angesprochen zumeist um ein zweistufiges Verfahren, bei dem innerhalb der Haushalte auf der zweiten Stufe eine Personenauswahl getroffen wird. Man kann sich hierbei leicht zusätzliche Verzerrungen einhandeln, wenn die Auswahl der Zielperson im Haushalt zum Beispiel nach dem Schwedenschlüssel oder mit der Geburtstagsmethode durch die Haushalte selbst vorgenommen wird (dies gilt allerdings nicht für CATI-Umfragen, bei denen die Auswahl der Zielperson normalerweise per Computer nach telefonischer Erfassung der Haushaltsstruktur erfolgt). Zudem führt die Zweistufigkeit des Verfahrens meistens zu einer Stichprobe mit nicht-identischen Auswahlwahrscheinlichkeiten, was tendenziell zu geringerer Stichprobeneffizienz führt und die Unannehmlichkeiten einer Gewichtung bei der Datenanalyse nach sich zieht. Weitere Probleme von Telefonstichproben betreffen zum Beispiel das Vorhandensein von Mehrfacheinträgen im Telefonregister beziehungsweise von unterschiedlichen Einträgen, die auf die gleichen Haushalte verweisen, und die damit verbundenen Schwierigkeiten, die Auswahlwahrscheinlichkeiten einzelner Personen genau zu bestimmen.

Alternativen zu den Stichproben aus dem Telefonregister gibt es wenige, so dass zum Beispiel auch das Schweizerische Bundesamt für Statistik für Großbefragungen wie etwa die jährlich durchgeführte Schweizerische Arbeitskräfteerhebung (SAKE) darauf zurückgreift. Auch stecken Random-Digit-Dialing-Verfahren (RDD), mit denen man zumindest die nicht eingetragenen Haushalte erreichen könnte, in der Schweiz noch in den Kinderschuhen. Erste Gehversuche werden zurzeit vom LINK-Institut unternommen und zudem scheint die BIK Asch-

purwis + Behrens GmbH aus Hamburg neuerdings RDD-Stichproben für die Schweiz anzubieten, die sich an das Verfahren von Gabler und Häder anlehnen (1997; vgl. auch den Übersichtsartikel von Häder und Glemser 2006). Ein Problem ist aber natürlich auch bei RDD-Verfahren, dass ein Teil der Bevölkerung so nicht erreichbar ist und der Schritt vom Telefonanschluss zur effektiv zu interviewenden Zielperson steinig sein kann.

Telefonstichproben (und RDD-Stichproben, wenn ein entsprechendes Verfahren einsatzbereit ist) mögen also zwar verhältnismäßig einfach und preiswert zu erlangen sein, sie sind aber aus den angesprochen Gründen methodisch nicht wirklich befriedigend. Eine bessere Datenquelle zur Ziehung von qualitativ hochwertigen Stichproben wären da ohne Zweifel die Einwohnerregister der Gemeinden. Die Register der Gemeinden, obwohl zurzeit noch in recht heterogener Form bezüglich der Art der Datenerfassung und der enthaltenen Information vorliegend, geben sicherlich das aktuellste und vollständigste Abbild der Bevölkerung der Schweiz.[2] Man beachte, dass die Stichprobenziehung über die Gemeinderegister unter Umständen sogar den Einbezug der Anstaltsbevölkerung in die Stichprobe erlaubt (vgl. zum Thema Schnell 1991). Nicht abgedeckt werden einzig einige Randgruppen wie illegal eingewanderte Personen, die sich der Registrierung durch die Behörden bewusst entziehen. Geringfügige Unschärfen ergeben sich zudem aufgrund von Mobilitätsprozessen. Insgesamt ist aber an den Einwohnerregisterdaten eigentlich nur problematisch, dass die Daten nicht in zentralisierter Form vorliegen. Entschließt man sich, eine Stichprobe mit Hilfe der Register zu ziehen, muss man in einem ersten Schritt also ein Verfahren zur Auswahl von Gemeinden anwenden. In einem zweiten Schritt wird dann mit den Gemeindeverwaltungen Kontakt aufgenommen, um die Zielpersonen der Stichprobe zu bestimmen.

Ein Gemeinde-Auswahlverfahren kann sein, mit Hilfe des vom Bundesamt für Statistik geführten Gemeindeverzeichnisses, das unter anderem Angaben zur Anzahl Einwohner in den Gemeinden enthält, eine künstliche Liste der Schweizer Bevölkerung zu erstellen, und aus dieser Liste mit einem Zufallsalgorithmus eine einfache Wahrscheinlichkeitsauswahl zu ziehen. Gemeinden, die mit mindestens einem Treffer in der fiktiven Stichprobe enthalten sind, werden dann zwecks Ziehung einer entsprechenden Anzahl realer Zielpersonen aus deren Einwohnerregister angeschrieben. Der Nachteil dieses Verfahrens ist, dass es zu einer Stichprobe mit sehr feiner geographischer Granularität führt. Das heißt, in einer so gezogenen Stichprobe sind sehr viele Gemeinden enthalten und in einem großen Teil dieser Gemeinden muss vielleicht nur gerade je ein Interview durchgeführt werden. Eine Folge davon ist, dass die Adressbeschaffung sehr zeitaufwändig

2 Auf Bundesebene gibt es konkrete Bestrebungen zur Harmonisierung der Einwohnerregister insbesondere auch zum Zwecke bevölkerungsstatistischer Erhebungen. So verabschiedete das Parlament kürzlich ein neues Registerharmonisierungsgesetz (RHG).

und auch kostspielig wird. Zudem entstehen etwa bei Face-to-Face-Interviews hohe Reisespesen.

Eine Lösung zur Verminderung der geographischen Streuung und somit der Reduktion der Kosten ist die Ziehung einer Klumpenstichprobe, bei der zum Beispiel Gemeinden mit einem PPS-Verfahren (*Probability Proportional to Size*) gezogen werden, und dann pro gezogene Gemeinde eine fixe Anzahl Interviews durchgeführt wird.[3] Die Anzahl in der Stichprobe enthaltener Gemeinden kann so im Vergleich zu einer einfachen Wahrscheinlichkeitsauswahl stark reduziert werden. Allerdings leidet aber auch die statistische Effizienz, das heißt, die Schätzung von Populationsparametern auf Grundlage solcher Klumpenstichproben ist unter Umständen mit einer deutlich erhöhten Unsicherheit behaftet.

Vor dem Hintergrund, dass eine einfache Wahrscheinlichkeitsauswahl zwar gute statistische Eigenschaften aufweist, aber nur aufwändig zu erlangen ist, und umgekehrt bei einer kostengünstigen Klumpenstichprobe die statistische Effizienz zu wünschen übrig lässt, haben Fritschi, Meyer und Schweizer (1976) einen Vorschlag für eine „geteilte" Zufallsstichprobe ausgearbeitet. Das Verfahren hat sich in Anlehnung an die Herkunft der Autoren als der „Berner Stichprobenplan" in der Literatur niedergeschlagen. Zwar wurde der Stichprobenplan in der Schweiz für eine Reihe von Studien verwendet (die prominenteste Arbeit, bei der auf das Verfahren zurückgegriffen wurde, ist die Armutsstudie von Leu et al. 1997; vgl. weiterhin z.B. Meyer et al. 1982 oder Wydler et al. 1996). Da es sich beim Berner Stichprobenplan aber um einen viel versprechenden Ansatz handelt, dessen Vorteile gegenüber den üblichen Telefonregisterstichproben meiner Meinung nach bisher zu wenig genutzt werden, möchte ich nachfolgend einige Überlegungen zu diesem Verfahren präsentieren.

Im nächsten Abschnitt wird das von Fritschi et al. (1976) vorgeschlagene Verfahren genauer beschrieben. Es wird unter anderem gezeigt, dass das Verfahren zu leicht verzerrten Stichproben führt, und es wird ein korrigiertes Verfahren entwickelt, das diese Verzerrungen ausgleicht. Der dritte Abschnitt befasst sich mit einer nahe liegenden Vereinfachung des Verfahrens und es werden die Eigenschaften des Berner Stichprobenplanes anhand von Simulationsergebnissen illustriert. Der vierte Abschnitt fasst die Ergebnisse zusammen.

3 Beziehungsweise in großen Gemeinden unter Umständen ein Vielfaches dieser fixen Anzahl. Bei sehr schiefen Verteilungen wie der Gemeindegrößenverteilung macht nur ein PPS-Verfahren *mit Zurücklegen* Sinn (einzelne Gemeinden hätten in einem Verfahren ohne Zurücklegen Auswahlwahrscheinlichkeit größer eins), so dass eine Gemeinde mehrmals in die PPS-Stichprobe gelangen kann. In solchen Gemeinden ist ein entsprechendes Vielfaches an Zielpersonen zu ziehen.

2 Das „geteilte“ Stichprobenverfahren von Fritschi, Meyer und Schweizer

Der Ausgangspunkt des Klumpungsverfahrens nach Fritschi et al. (1976) ist die Überlegung, dass eine Klumpung der Stichprobe für relativ große Gemeinden wenig Sinn macht, weil auf diese Gemeinden auch bei einer einfachen Wahrscheinlichkeitsauswahl eine hinreichend große Anzahl Stichprobenmitglieder entfällt. Durch die Klumpung würde man in diesem Teil der Population unnötigerweise eine Menge Präzision verschenken. Das Problem sind vielmehr die vielen kleinen Gemeinden, in denen bei einer einfachen Wahrscheinlichkeitsauswahl jeweils nur eine minimale Anzahl Interviews durchgeführt werden muss. Fritschi et al. (1976) schlagen deshalb das folgende Verfahren zur Erstellung einer „geteilten“ Zufallsstichprobe vor: In einem ersten Schritt ziehe man ausgehend von einem Gemeindeverzeichnis mit Angaben zu den Einwohnerzahlen (bzw. der Anzahl Stimm- und Wahlberechtigter im Anwendungsfall von Fritschi et al.) eine (hypothetische) Stichprobe aus der durch das Gemeindeverzeichnis definierten Population. Die Ziehung erfolgt nach den Regeln einer einfachen Wahrscheinlichkeitsauswahl mit konstanter Auswahlwahrscheinlichkeit $p = n / N$, wobei n dem gewünschten Umfang der Stichprobe und N der Populationsgröße entspricht. Für jede Gemeinde wird dann die Anzahl „Treffer“, also die Anzahl hypothetischer Stichprobenmitglieder, die auf diese Gemeinde entfallen sind, gezählt. Aus der durch die Gemeinden mit weniger als k Treffern definierten Teilpopulation wird dann in einem zweiten Schritt eine Stichprobe mit reduzierter Trefferwahrscheinlichkeit p / k gezogen. Die Endstichprobe setzt sich schließlich zusammen aus einer der Anzahl Treffer entsprechen Anzahl Zielpersonen in den Gemeinden mit k oder mehr Treffern im ersten Schritt und jeweils k Zielpersonen in den Gemeinden mit einem Treffer im zweiten Schritt (bzw. ein der Anzahl Treffer entsprechendes Vielfaches von k, wenn eine Gemeinde im zweiten Schritt mehr als einen Treffer verzeichnet). Die Bestimmung der effektiven Zielpersonen erfolgt dann in der Praxis am besten über die Einwohnerregister der in der Stichprobe enthaltenen Gemeinden. Alternativ wäre auch ein Random-Route-Verfahren denkbar.

Der Ansatz von Fritschi et al. (1976) ist einleuchtend: Zuerst eine einfache Wahrscheinlichkeitsauswahl durchführen, dann denjenigen Teil der Stichprobe, in dem nur wenige Treffer pro Gemeinde vorliegen, durch eine Klumpenstichprobe ersetzen. Anders als von Fritschi et al. vermeintlich bewiesen, handelt es sich aber nicht um eine Methode, bei der die *a priori* Auswahlwahrscheinlichkeiten in allen Gemeinden gleich sind, und die Verteilung von mit der Gemeindegröße zusammenhängenden Merkmalen wird verzerrt wiedergegeben. Dies soll nun dargelegt werden.

Sei P_{ij} die Wahrscheinlichkeit, dass Person i aus Gemeinde j in die Stichprobe gelangt. Bei einer einfachen Wahrscheinlichkeitsauswahl ist P_{ij} für alle Personen aus der Grundgesamtheit gleich, also

$$P_{ij} = p = \frac{n}{N} \text{ für alle } i, j$$

wenn n der Stichprobenumfang und N die Populationsgröße ist. Stichproben, die diese Eigenschaft besitzen, werden manchmal als EPSEM-Stichproben bezeichnet (*Equal Probability of Selection Method*, Babbie 1979, S. 330). Zwar ist es möglich, auch aus Stichproben mit nicht-identischen Auswahlwahrscheinlichkeiten auf die Grundgesamtheit zu schließen, sofern die Auswahlwahrscheinlichkeiten bekannt sind. Im Allgemeinen sind jedoch EPSEM-Stichproben aus theoretischen (mehr statistische Effizienz) sowie praktischen Gründen (es werden weniger komplexe Schätzer benötigt) vorzuziehen (eine Ausnahme mögen je nach Eigenschaften des Untersuchungsgegenstandes disproportional geschichtete Stichproben bilden). Im vorliegenden Fall eines gemeindebasierten Stichprobenverfahrens sollte insbesondere darauf geachtet werden, dass die Auswahlwahrscheinlichkeiten nicht mit der Gemeindegröße zusammenhängen (oder aber, dass der Zusammenhang genau bekannt ist und mit entsprechenden Gewichten bei der Datenanalyse ausgeglichen werden kann).

Wie erläutert, wird beim zweistufigen Berner Stichprobenplan in einem ersten Schritt eine einfache Wahrscheinlichkeitsauswahl mit Auswahlwahrscheinlichkeit $p = n / N$ durchgeführt. Es wird sodann die Anzahl „Treffer", also die Anzahl ausgewählter Personen pro Gemeinde gezählt und in einem zweiten Schritt unter den Gemeinden, die eine kritische Anzahl Treffer k nicht erreicht haben, eine Klumpenstichprobe gezogen. Die Stichprobe setzt sich schließlich zusammen aus den im ersten Schritt gewählten Personen in den Gemeinden mit mehr als k Treffern und den Klumpen zu je k Personen aus dem zweiten Schritt. Sei X eine Zufallsvariable der Anzahl Treffer pro Gemeinde im ersten Schritt und Y eine von X unabhängige Zufallsvariable der Anzahl Klumpen pro Gemeinde im zweiten Schritt. Die Auswahlwahrscheinlichkeit von Person i aus Gemeinde j kann dann geschrieben werden als

$$P_{ij} = P\left(X_j \geq k\right) \frac{E\left(X_j \middle| X_j \geq k\right)}{N_j} + P\left(X_j < k\right) \frac{E\left(Y_j \middle| X_j < k\right) \cdot k}{N_j} \qquad (1)$$

wobei $P(X_j \geq k) = 1 - P(X_j < k)$ der Wahrscheinlichkeit entspricht, dass in Gemeinde *j* im ersten Schritt *k* oder mehr Treffer erreicht werden. N_j entspricht der Anzahl Personen in Gemeinde *j* und $E(.)$ symbolisiert den Erwartungswert.

Wird mit Zurücklegen gezogen, können *X* und *Y* als binomialverteilte Zufallsvariablen aufgefasst werden, also

$$X_j \sim B(N_j, p) \text{ und } Y_j \sim B(N_j, \pi_j)$$

mit *p* und π_j als den Trefferwahrscheinlichkeiten. Es gilt insbesondere

$$E(X_j) = p \cdot N_j \text{ und } E(Y_j) = \pi_j N_j$$

Aus der Unabhängigkeit von *X* und *Y* folgt zudem $E(Y_j | X_j < k) = E(Y_j)$. Wenn der Berner Stichprobenplan nun EPSEM-Charakter haben soll, dann muss $P_{ij} = p = n / N$ gelten (für alle *i* und *j*). Durch Einsetzen in (1) erhält man nach einigen Umformungen jedoch

$$\pi_j = \frac{\frac{n}{N} - P(X_j \geq k) \frac{E(X_j | X_j \geq k)}{N_j}}{k \cdot [1 - P(X_j \geq k)]} \qquad (2)$$

Auch wenn $P(X_j \geq k)$ und $E(X_j | X_j > k)$ weiter aufgelöst werden (siehe Anhang), lässt sich N_j nicht aus Gleichung (2) eliminieren. Das heißt, π_j hängt von der Gemeindegröße ab und ist nicht konstant, was in Widerspruch zu den Ausführungen von Fritschi et al. (1976) steht, die von einer konstanten Auswahlwahrscheinlichkeit $\pi = p / k$ ausgehen. Es ist also zu erwarten, dass der Berner Stichprobenplan nach Fritschi et al., der im ersten wie auch im zweiten Schritt jeweils eine konstante Auswahlwahrscheinlichkeit vorsieht, zu verzerrten Ergebnissen führt. Die Verzerrung kann indes korrigiert werden, indem im zweiten Schritt variable Auswahlwahrscheinlichkeiten nach Formel (2) eingesetzt werden.[4]

Die Eigenschaften der Verzerrung im unkorrigierten Berner Stichprobenplan können mit Hilfe einer Simulation veranschaulicht werden. Als Grundlage für die Simulation verwende ich das Gemeindeverzeichnis 2002 der Schweiz, das Anga-

4 Ein alternatives Korrekturverfahren mit einem ähnlichen Effekt bestünde darin, die Auswahlwahrscheinlichkeiten im ersten Schritt variabel zu gestalten, dafür aber im zweiten Schritt konstant zu halten (die erforderlichen Größen lassen sich ebenfalls aus den hier präsentierten Formeln ableiten). Eine weitere Korrekturstrategie wäre, die Stichprobe mit dem unkorrigierten Verfahren zu ziehen, jedoch bei der Datenanalyse aus Formel (1) ableitbare Design-Gewichte einzusetzen.

ben zu den Einwohnerzahlen aller politischen Gemeinden enthält (Bundesamt für Statistik 2002). Gezogen wird jeweils eine Stichprobe im Umfang von n=3.750 Personen, die minimale Anzahl Zielpersonen pro Gemeinde ist auf k=10 festgelegt (dies entspricht bezüglich der Anzahl in die Stichprobe gelangender Gemeinden in etwa den im Anwendungsbeispiel von Fritschi et al. verwendeten Parametern; Fritschi et al. legten ihrem Verfahren die etwas kleinere Population der Wahlberechtigten zu Grunde). Tabelle 1 zeigt die Simulationsergebnisse über 10'000 Runden. Angegeben ist einerseits die faktische Gemeindegrößenverteilung in der Population und andererseits die aus den Stichproben resultierende gemittelte Gemeindegrößenverteilung für verschiedene Ziehungsverfahren. Man erkennt, dass bei der ursprünglichen Version des Berner Stichprobenplans die meisten Gemeindegrößenkategorien leicht unterrepräsentiert sind. Zum Beispiel leben 12.06 Prozent der Bevölkerung in Gemeinden mit 2000 bis 3499 Einwohnern, in den Stichproben auf Grundlage des Berner Stichprobenplans sind es im Schnitt aber nur 11.71 Prozent. Eine Kategorie jedoch, Gemeindegröße 10'000 bis 49'999, ist deutlich übervertreten. Ähnlich wie bei einer einfachen Wahrscheinlichkeitsauswahl (letzte Spalte) treten diese Verzerrungen beim „korrigierten" Berner Stichprobenplan, bei dem die Auswahlwahrscheinlichkeiten im zweiten Schritt nach Formel (2) bestimmt wurden, jedoch nicht auf.

Tabelle 1: Gemeindegrößenverteilung im Berner Stichprobenplan

Gemeindegröße	Bevölkerungs-anteil	Stichprobenverteilungen (10'000 Replikationen)		
		Berner Stichprobenplan (unkorrigiert)	Berner Stichprobenplan (korrigiert)	einfache Zufallsauswahl
1 – 249	0.98	0.94	0.98	0.98
250 – 499	2.47	2.39	2.48	2.47
500 – 999	5.56	5.36	5.58	5.57
1000 – 1999	10.49	10.18	10.48	10.48
2000 – 3499	12.06	11.71	12.08	12.07
3500 – 4999	9.89	9.60	9.87	9.90
5000 – 9999	16.79	16.49	16.77	16.80
10000 – 49999	25.67	27.25	25.66	25.65
50000 und mehr	16.09	16.08	16.10	16.08
Total	100.00	100.00	100.00	100.00

Quelle: Simulationsergebnisse (10'000 Replikationen) auf Grundlage des Gemeindeverzeichnisses 2002 der Schweiz (BFS 2002, 2876 Gemeinden mit insgesamt 7'241'468 Einwohnern) mit Ziel-Stichprobengröße 3750 und Klumpengröße 10.

Ein eindrückliches Bild der Natur der Verzerrung im unkorrigierten Berner Stichprobenplan liefert Abbildung 1. Dargestellt ist die relative Verteilung der gemittelten Gemeindegrößenverteilung in den Stichproben und der Gemeindegrößenverteilung in der Population (bzw. genauer: für jede Gemeinde ist das Verhältnis ihres durchschnittlichen Anteils an der Stichprobe und ihres tatsächlichen Anteils an der Gesamtbevölkerung abgebildet). Man sieht sehr deutlich, dass Personen aus Gemeinden mit ca. 10'000 bis 30'000 Einwohnern zu häufig in die Stichprobe gelangen. Natürlich sind dies gerade diejenigen Gemeinden, die an der „Grenze" zur Klumpenstichprobe liegen. Das sind Gemeinden mit einer durchschnittlich erwarteten Trefferzahl nahe der Mindestklumpengröße k, die also je nach Ausgang der einfachen Wahrscheinlichkeitsauswahl im ersten Schritt manchmal in das Klumpenauswahlverfahren im zweiten Schritt gelangen und manchmal nicht. Beim korrigierten Berner Stichprobenplan mit Auswahlwahrscheinlichkeiten nach Formel (2) wird die Verzerrung, wie man erkennen kann, erfolgreich geglättet.[5]

5 Zur Verteidigung von Fritschi et al. (1976) ist anzumerken, dass die hier recht bedeutend erscheinende Verzerrung in der von Fritschi et al. gewählten praktischen Umsetzung ihres Stichprobenverfahrens deutlich geringer war. Fritschi et al. verwendeten den damals zur Verfügung stehenden technischen Möglichkeiten entsprechend einen Ziehungsalgorithmus, bei dem ausgehend von einem Zufallsstartwert die nach Gemeinden strukturierte Population mit fixen Auswahlschritten durchwandert wird. Dieses Verfahren führt zu einer geringeren Variation der Anzahl Treffer in einer Gemeinde mit gegebener Gemeindegröße als ein „echtes" Zufallsverfahren, was die Verzerrung der durch den Stichprobenplan erzeugten Gemeindegrößenverteilung fast vollständig eliminiert. Der von Fritschi et al. angewendete Algorithmus kann als Verfahren angesehen werden, bei dem quasi nach Gemeindegröße geschichtet gezogen wird. Eine systematische Schichtung nach Gemeindegröße kann durchaus wünschenswert sein. Allerdings sollte man dann gleich ein „echtes" Schichtungsverfahren verwenden, dessen theoretische Eigenschaften bekannt sind.

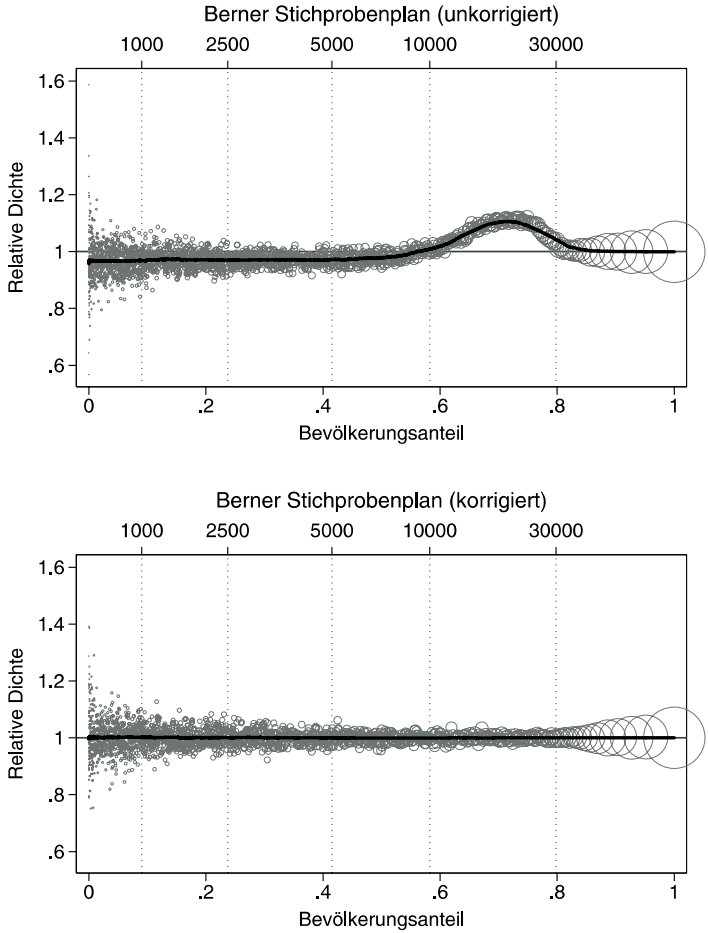


Abb.1: Vergleich der Gemeindeverteilung des Berner Stichprobenplans (gemittelt über 10'000 Ziehungen) mit der Gemeindeverteilung in der Grundgesamtheit

Legende: Dargestellt ist für jede Gemeinde das Verhältnis zwischen ihrem durchschnittlichen Anteil in der Stichprobe (10'000 Ziehungen mit Stichprobengröße 3750 und Klumpengröße 10) und ihrem Anteil an der Gesamtbevölkerung. Ein Wert von eins bedeutet, dass die Mitglieder der Gemeinde im Durchschnitt genau ihrem Anteil an der Gesamtbevölkerung entsprechend repräsentiert werden (zum Konzept der relativen Verteilung siehe Handcock und Morris 1999). Die Gemeinden sind ihrer Größe nach auf der proportional zum kumulierten Bevölkerungsanteil skalierten X-Achse geordnet, zudem richtet sich die Größe der Symbole nach der Anzahl Einwohner in den entsprechenden Gemeinden. Da in einer endlichen Simulation die Werte vor allem für klei-

ne Gemeinden beträchtlich streuen, wurde zur Verdeutlichung des Trends eine geglättete Kurve eingezeichnet (Sasieni 1998).

3 Ein vereinfachtes „geteiltes" Verfahren

Zwar ist es, wie oben dargelegt, möglich, das Verfahren von Fritschi et al. (1976) so zu modifizieren, dass unverzerrte Stichproben erzeugt werden. Alternativ könnten auch nachträglich Gewichte berechnet werden, um der Verzerrung bei der Datenanalyse entgegenzuwirken. Ich möchte hier aber zeigen, dass beide Möglichkeiten wenig praktische Relevanz besitzen, da es ein Verfahren zur Ziehung einer „geteilten" Stichprobe gibt, das in vergleichbarer Art zu einer Reduktion der Anzahl Gemeinden in der Stichprobe führt, jedoch einfacher umzusetzen ist und erst noch etwas bessere statistische Eigenschaften aufweist. Die Komplikationen des Verfahrens von Fritschi et al. rühren daher, dass die Gemeinden *ex post*, also erst *nach* der ersten Ziehung einer einfachen Wahrscheinlichkeitsauswahl aufgrund der empirisch realisierten Anzahl Treffer in zwei Gruppen eingeteilt werden. Eine offensichtliche Strategie zur Vermeidung dieser Probleme ist, die Gruppenzuteilung ganz einfach *ex ante* aufgrund der theoretisch erwarteten Anzahl Treffer vorzunehmen. Ich schlage also ein Verfahren vor, bei dem – gegeben die Trefferwahrscheinlichkeit $p = n / N$ – die Gemeinden zuerst in eine Gruppe mit $E(X_j) = p \cdot N_j \geq k$ und eine Gruppe mit $E(X_j) < k$ aufgeteilt werden. In einem zweiten Schritt wird dann aus der in Gruppe 1 mit den großen Gemeinden enthaltenen Population eine einfache Wahrscheinlichkeitsauswahl gezogen und aus Gruppe 2 mit den kleinen Gemeinden eine Klumpenstrichprobe. Der angestrebte Stichprobenumfang n wird dabei im Verhältnis der durch die beiden Gruppen abgedeckten Bevölkerungsanteile aufgeteilt. Das heißt, der zu realisierende Stichprobenanteil in Gruppe 1 ist

$$n_1 = \frac{n}{N} \cdot \sum_{E(X_j) \geq k} N_j = \sum_{E(X_j) \geq k} E(X_j)$$

und die Anzahl in Gruppe 2 zu ziehender Klumpen beträgt

$$n_k = \frac{n_2}{k} = \frac{n}{k \cdot N} \cdot \sum_{E(X_j) < k} N_j = \frac{1}{k} \sum_{E(X_j) < k} E(X_j)$$

wobei bei nicht ganzzahligen Ergebnissen für n_1 und n_k in der Praxis Zufallsrundungen einzusetzen sind. Es folgt unmittelbar, dass es sich bei der beschriebenen

ex ante geteilten Zufallsstichprobe um ein Verfahren mit identischen Auswahlwahrscheinlichkeiten handelt, da

$$P_{ij|E(X_j)\geq k} = \frac{n_1}{\sum_{E(X_j)\geq k} N_j} = P_{ij|E(X_j)<k} = \frac{k \cdot n_k}{\sum_{E(X_j)<k} N_j} = \frac{n}{N}$$

Die praktische Durchführung der Ziehung einer *ex ante* geteilten Zufallsstichprobe ist bedeutend einfacher als die Ziehung einer Stichprobe nach dem korrigierten Verfahren von Fritschi et al. (1976), da keine komplizierten Ergebnisse wie Formel (2) benötigt werden. Das Verfahren hat aber noch einen anderen Vorteil: Durch die vorgängige Aufteilung der Population entsteht ein Schichtungseffekt, der die Effizienz der Stichprobe für Merkmale, die einen Zusammenhang zur Gemeindegröße aufweisen, im Vergleich zum Stichprobenplan von Fritschi et al. zusätzlich erhöht.

Tabelle 2 zeigt die Ergebnisse einer Simulation, bei der das hier vorgeschlagene *ex ante* geteilte Stichprobenverfahren mit einer einfachen Wahrscheinlichkeitsauswahl, einer Klumpenstichprobe und dem *ex post* geteilten Stichprobenverfahren (dem korrigierten Berner Stichprobenplan nach Formel 2) verglichen wird. Grundlage der Simulation ist wiederum das Gemeindeverzeichnis 2002 der Schweiz (Bundesamt für Statistik 2002). Gezogen werden Stichproben im Umfang von n=1000. Die Klumpengröße in den geklumpten Stichproben beträgt k=8.[6] Die Simulationsergebnisse in Tabelle 2 beziehen sich auf die Stichprobenschätzer der Erwartungswerte von verschiedenen Merkmalen mit unterschiedlichem linearen Zusammenhang zu der Gemeindegröße (r) und mit unterschiedlicher Intra-Klassen-Korrelation in den Gemeinden (ρ; die Intra-Klassen-Korrelation ist ein Maß für die interne Homogenität in den Gemeinden). Sämtliche Merkmale sind standardisiert, das heißt, sie haben in der Population einen Mittelwert von 0 und eine Standardabweichung von 1. Dargestellt sind für jedes Stichprobenverfahren die Durchschnitte und Standardabweichungen der Mittelwertsschätzer über 10'000 Ziehungen.

6 Man beachte, dass mit diesen Simulationsparametern nur gerade acht der knapp 3000 Gemeinden in den ungeklumpten Zweig der *ex ante* geteilten Stichprobe gelangen (es sind dies Zürich, Genf, Basel, Bern, Lausanne, Winterthur, St. Gallen und Luzern).

Tabelle 2: Mittelwerte und Standardfehler (in Klammern) von Erwartungswertsschätzern verschiedener Stichprobenverfahren

		Simulationsergebnisse (10'000 Replikationen)			
	Erwartungswert	einfache Zufallsauswahl	Klumpenstichprobe	Berner Stichprobenplan	
				ex post geteilte Stichprobe	*ex ante* geteilte Stichprobe
X1 ($r=1.00, \rho=1.00$)	0.000	0.000	0.001	0.000	0.000
		(0.032)	(0.090)	(0.033)	(0.019)
X2 ($r=0.50, \rho=0.25$)	0.000	0.000	0.000	0.000	0.000
		(0.031)	(0.052)	(0.032)	(0.029)
X3 ($r=0.50, \rho=0.50$)	0.000	0.000	0.000	0.000	0.001
		(0.031)	(0.067)	(0.051)	(0.049)
X4 ($r=0.00, \rho=0.50$)	0.000	0.000	0.001	0.000	0.001
		(0.031)	(0.067)	(0.064)	(0.065)
X5 ($r=0.00, \rho=0.50^a$)	0.000	0.000	0.001	0.000	0.000
		(0.032)	(0.068)	(0.033)	(0.032)
X6 ($r=0.00, \rho=0.00$)	0.000	0.000	0.000	0.000	0.000
		(0.032)	(0.032)	(0.032)	(0.032)
Anzahl PSU		1000.0	125.0	275.1	265.8
Anzahl Gemeinden		568.8	104.7	105.9	105.9
Fallzahl		1000.0	1000.0	1000.0	1000.0

Quelle: Simulationsergebnisse (10'000 Replikationen) auf Grundlage des Gemeindeverzeichnisses 2002 der Schweiz (BFS 2002, 2876 Gemeinden mit insgesamt 7'241'468 Einwohnern) mit Ziel-Stichprobengröße 1000 und Klumpengröße 8. Bei den Variablen X1 bis X6 handelt es sich um künstlich für die Population erzeugte Merkmale (r: Korrelation mit der Gemeindegröße; ρ: Intra-Klassen-Korrelation, [a] mit $\rho=1$ in den großen und $\rho=0$ in den kleinen Gemeinden).

Im Fuß von Tabelle 2 finden sich einige Angaben zur durchschnittlichen Anzahl PSU (*Primary Sampling Units*, d.h. Anzahl Klumpen)[7] und der durchschnittlichen Anzahl Gemeinden in den Stichproben. Man erkennt sehr schön, dass alle Klumpungsverfahren in vergleichbarer Weise zu einer Reduktion der Anzahl Gemeinden auf etwas mehr als 100 führen – im Vergleich zu den knapp 600 Gemein-

7 Bei der einfachen Wahrscheinlichkeitsauswahl sind die PSU identisch mit den Zielpersonen, das heisst, eine Stichprobe von Umfang n enthält n Klumpen mit jeweils einem Mitglied. Die reine Klumpenstichprobe enthält n/k Klumpen mit je k Mitgliedern. Bei den geteilten Verfahren enthalten die Stichproben eine Mischung aus Ein-Personen- und k-Personen-Klumpen.

den, die bei einer einfachen Wahrscheinlichkeitsauswahl in die Stichprobe gelangen. Verbunden mit der Verringerung der Anzahl Gemeinden ist naturgemäß eine Reduktion der Anzahl PSU, wobei aber diese Reduktion beim Berner Stichprobenplan (275 bzw. 266 PSU) deutlich geringer ausfällt als bei der reinen Klumpenstichprobe (125 PSU). Gerade in dieser geringeren Reduktion der Anzahl PSU bei gleichzeitig vergleichbarer Reduktion der Anzahl Gemeinden liegt die Stärke des Berner Stichprobenplans gegenüber der reinen Klumpenstichprobe, wobei dieser Vorteil allerdings nur unter bestimmten, noch zu identifizierenden Bedingungen tatsächlich zu einer verbesserten Stichprobeneffizienz führt.

An den Resultaten im Hauptteil von Tabelle 2 erkennt man, dass mit allen Stichprobenverfahren die Populationsmittelwerte der vier Merkmale erwartungstreu geschätzt werden: Die Populationsmittelwerte werden mit allen Verfahren im Mittel über die 10'000 Replikationen praktisch exakt reproduziert. Die Verfahren unterscheiden sich aber deutlich hinsichtlich der Streuung der Mittelwertsschätzer über die einzelnen Stichproben. Zum Beispiel ist bei der Klumpenstichprobe der Standardfehler des Mittelwertschätzers für Merkmal X1, welches perfekt mit der Gemeindegröße korreliert (X1 ist eine lineare Transformation der Gemeindegröße), fast dreimal so groß wie bei der einfachen Wahrscheinlichkeitsauswahl. Der Design-Effekt in der Simulation, also das Verhältnis der Varianzen, beträgt in diesem Fall 8.12, was in etwa dem nach der Formel von Kish (1965, S. 162) erwarteten Design-Effekt für Klumpenstichproben von $1+\rho(k-1)=8$ entspricht (vgl. Tabelle 3). Beim (korrigierten) Berner Stichprobenplan ist der Standardfehler für Merkmal X1 jedoch nur geringfügig größer als bei der einfachen Zufallsauswahl und beim *ex ante* geteilten Verfahren ist er sogar bedeutend kleiner (der Design-Effekt beträgt 0.36). Letzteres liegt am angesprochenen Schichtungseffekt durch die vorgängige Teilung der Stichprobe in zwei Gemeindegrößengruppen, aus denen dann separate Stichproben gezogen werden.

Bei Merkmal X2, das mittelstark mit der Gemeindegröße korreliert, aber über keine über das durch den linearen Zusammenhang bedingte Ausmaß hinausgehende Intra-Klassen-Korrelation verfügt,[8] schneidet der Berner Stichprobenplan ebenfalls sehr gut ab. Der Design-Effekt ist für die *ex ante* geteilte Stichprobe immer noch Varianz vermindernd und beträgt 0.86 (gegenüber 2.78 für die Klumpenstichprobe; Tabelle 3, Spalte 2). Wird nun die Intra-Klassen-Korrelation aber systematisch erhöht, verlieren die geteilten Stichprobenverfahren zunehmend an Boden. Merkmal X3 weist wie X2 einen linearen Zusammenhang von 0.5 auf, die Intra-Klassen-Korrelation wurde aber künstlich von den minimal 0.25 auf ebenfalls 0.5 angehoben. Die Streuung ist nun in der *ex ante* geteilte Stichprobe deutlich höher als in der einfachen Wahrscheinlichkeitsauswahl, der Design-Effekt ist

8 Aus dem linearen Zusammenhang zwischen einem Merkmal X und der Gemeindegröße folgt ein Mindestmaß an interner Homogenität in den Gemeinden von $p^{min} = r^2$.

aber mit 2.44 immer noch nur etwa halb so groß wie bei der reinen Klumpenstichprobe.

Tabelle 3: Simulierte und theoretische Design-Effekte (Verhältnis der Varianzen)

	X1 (r=1.00, ρ=1.00)	X2 (r=0.50, ρ=0.25)	X3 (r=0.00, ρ=0.50)	X4 (r=0.00, ρ=0.50)	X5 (r=0.00, ρ=0.50[a])	X6 (r=0.00, ρ=0.00)
Klumpenstichprobe vs. einfache Zufallsauswahl	8.12	2.78	4.61	4.60	4.57	1.00
ex ante geteilte Stichprobe vs. einfache Zufallsauswahl	0.36	0.86	2.44	4.22	1.00	1.01
ex ante geteilte Stichprobe vs. *ex post* geteilte Stichprobe	0.33	0.83	0.92	1.00	0.93	1.01
Theoretischer Design-Effekt für eine Klumpenstichprobe mit 125 Klumpen	8.00	2.75	4.50	4.50	4.50	1.00
Theoretischer Design-Effekt für eine Klumpenstichprobe mit 266 Klumpen	3.76	1.69	2.38	2.38	2.38	1.00

Quelle: Beruhend auf den Zahlen in Tabelle 2. r: Korrelation mit der Gemeindegröße; ρ: Intra-Klassen-Korrelation, [a] mit $\rho=1$ in den großen und $\rho=0$ in den kleinen Gemeinden

Die Vorteile des Berner Stichprobenplans verschwinden schließlich fast vollständig, wenn die interne Homogenität der Gemeinden auf hohem Niveau beibehalten, der Zusammenhang zur Gemeindegröße aber eliminiert wird. Merkmal X4 ist unabhängig von der Gemeindegröße an sich, weist aber eine relativ starke interne Homogenität in den Gemeinden auf. Das heißt, Personen aus der gleichen Gemeinde sind sich bezüglich X4 – unabhängig von der Gemeindegröße – im Durchschnitt bedeutend ähnlicher als Personen aus unterschiedlichen Gemeinden. Die Eigenschaften des Berner Stichprobenplans verschlechtern sich hier verglichen mit der einfachen Wahrscheinlichkeitsauswahl deutlich. Der Design-Effekt erreicht nun fast einen Wert, wie er für die reine Klumpenstichprobe festgestellt werden kann, und liegt – etwas überraschend – deutlich über dem theoretisch erwarteten Design-Effekt für eine reine Klumpenstichprobe mit einer dem Wert für die geteilte Stichprobe entsprechenden Anzahl von 266 Klumpen bzw. PSU (vgl. die letzte Zeile von Tabelle 3). Das heißt, dass sich die größere Anzahl PSU im geteilten Verfahren bei Merkmal X4 nicht voll ausspielen kann. Daraus lässt sich schließen, dass es auf die genaue Struktur ankommt, in der sich die Intra-Klassen-Korrelation eines Merkmals präsentiert. Bei den bisherigen Merkmalen wurde von homoskedastischer Varianz innerhalb der Gemeinden ausgegan-

gen. Die Intra-Klassen-Korrelation kann nun aber wiederum selbst eine Funktion der Gemeindegröße sein, was sich offensichtlich auf die Effizienz des Berner Stichprobenplans auszuwirken scheint. Merkmal X5 repräsentiert den Extremfall, in dem in den großen Gemeinden mit $E(X_j) \geq k$ perfekte interne Homogenität herrscht, die kleinen Gemeinden mit $E(X_j) < k$ jedoch eine Intra-Klassen-Korrelation von $\rho = 0$ aufweisen. Die Varianzen in den beiden Teilen der Population wurden zudem so gewählt, dass sich über alle Gemeinden hinweg eine Intra-Klassen-Korrelation von 0.5 ergibt. An den Simulationsresultaten in den Tabellen 2 und 3 erkennt man, dass in dieser Situation, von der man auf den ersten Blick annehmen könnte, sie unterscheide sich kaum von der Situation für Merkmal X4, das *ex ante* geteilte Stichprobenverfahren wiederum sehr gut abschneidet und einen neutralen Design-Effekt von 1 erreicht. An was liegt das? Die Begründung ist klar: In den kleinen Gemeinden ist die geteilte Stichprobe eine Klumpenstichprobe und bei einer Intra-Klassen-Korrelation von 0 ist die Klumpenstichprobe gleich effizient wie eine einfache Wahrscheinlichkeitsauswahl (vgl. unten). In den großen Gemeinden entspricht die geteilte Stichprobe einer einfachen Wahrscheinlichkeitsauswahl, die bei einer Intra-Klassen-Korrelation von 1 einen maximalen Effizienzvorteil gegenüber der Klumpenstichprobe aufweist. Die *ex ante* geteilte Stichprobe *muss* in dieser Situation folglich einen Design-Effekt von 1 haben.[9] Umgekehrt lässt sich auch ableiten, unter welchen Bedingungen die geteilte Stichprobe genau gleich ineffizient ist wie eine reine Klumpenstichprobe. Dies ist der Fall, wenn in den kleinen Gemeinden perfekte interne Homogenität herrscht und die großen Gemeinden perfekte Heterogenität aufweisen. Gegeben die Korrelation zwischen einem Merkmal und der Gemeindegröße ist null, liegt der Berner Stichprobenplan bezüglich der statistischen Effizienz also je nach Struktur der internen Homogenität in den Gemeinden zwischen den beiden Polen der einfachen Wahrscheinlichkeitsauswahl und der Klumpenstichprobe. Bei Korrelation zwischen dem Merkmal und der Gemeindegröße (bzw. genauer: bei positiver Intra-Klassen-Korrelation in den beiden Gemeindegrößenschichten) kommt beim *ex ante* geteilten Stichprobenverfahren zusätzlich der Schichtungseffekt zum Tragen, so dass für einzelne Merkmale unter Umständen eine größere statistische Effizienz als in der einfachen Wahrscheinlichkeitsauswahl erreicht werden kann.

Merkmal X6 illustriert weiterhin den Fall, in dem mit allen Stichprobenverfahren die gleichen Ergebnisse erzielt werden. Wenn keine Intra-Klassen-Korrelation in den Gemeinden besteht, das heißt, wenn sich Personen aus der gleichen Gemeinde bezüglich eines Merkmals nicht ähnlicher sind als Personen aus unterschiedlichen Gemeinden, dann spielt es für die Effizienz der Stichprobe keine

9 Die Tatsache, dass die *ex ante* geteilte Stichprobe geschichtet ist, spielt hier keine Rolle, da X4 unkorreliert ist mit der Gemeindegrösse. Das heisst, der Mittelwert von X4 ist in beiden Schichten identisch und es kann somit kein Schichtungseffekt entstehen.

Rolle, ob die Stichprobe nach Gemeinden geklumpt ist oder nicht. Selbst in der reinen Klumpenstichprobe ist in diesem (in der Realität wohl eher selten anzutreffenden) Fall also der Standardfehler nicht größer als in der einfachen Wahrscheinlichkeitsauswahl.

Tabelle 3 mit den Design-Effekten enthält schließlich noch den Vergleich zwischen dem *ex ante* geteilten Stichprobenverfahren und dem *ex post* geteilten Verfahren, bei dem die Auswahlwahrscheinlichkeiten nach Formel (2) bestimmt wurden. Man erkennt, dass aufgrund des Schichtungseffekts die *ex ante* geteilte Stichprobe in fast allen Situationen zumindest geringfügig besser abschneidet als die *ex post* geteilte Stichprobe und nie eine größere Varianz aufweist. Die Wahl des einfacher umzusetzenden *ex ante* geteilten Stichprobenverfahrens ist somit die „dominante" Strategie: man kann sich gegenüber dem *ex post* geteilten Verfahren nur verbessern, nicht aber verschlechtern.

4 Zusammenfassung und Diskussion

Die in der Schweiz am häufigsten eingesetzte Methode zur Erstellung einer Bevölkerungsstichprobe – die Ziehung aus dem Telefonregister – ist vor allem aufgrund mangelnder Abdeckung der Grundgesamtheit nur mit Vorbehalt für Studien, die wissenschaftliche Qualitätsstandards erfüllen sollen, in Betracht zu ziehen. Der Berner Stichprobenplan, bei dem die Stichprobe über die Einwohnerregister der Gemeinden bestimmt wird, kann hier eine Alternative sein. Zwar sind die Kosten und der benötigte Zeitaufwand der Adressbeschaffung höher als bei einer Telefonregisterstichprobe, die zweifelsohne stark verbesserte Qualität der resultierenden Ausgangsstichprobe lässt den Zusatzaufwand, der sich durch die Klumpung der Stichprobe auf wenige Gemeinden in überschaubarem Rahmen bewegt, aber durchaus als lohnenswert erscheinen. Ohnehin ist der Aufwand für die Stichprobenziehung eher verschwindend, wenn er am Aufwand gemessen wird, der dann in der Regel für die Durchführung der Interviews zu betreiben ist. Man spart sicherlich am falschen Ort, wenn man auf billige, qualitativ schlechte Stichprobenverfahren setzt, auch wenn durch die Kosteneinsparung ein paar zusätzliche Interviews realisiert werden können. Denn was nützt eine größere Fallzahl, wenn die Stichprobe verzerrt ist?

Die Eigenschaften des Berner Stichprobenplans wurden in den Abschnitten 2 und 3 erläutert. Es wurde erstens gezeigt, dass das ursprünglich von Fritschi et al. (1976) vorgeschlagene Verfahren zu leicht verzerrten Stichproben führt, was aber mit ein paar Modifikationen korrigiert werden kann. Zudem wurde mit dem *ex ante* geteilten Stichprobenverfahren ein alternativer Ansatz vorgeschlagen, der einfacher umzusetzen ist und erst noch etwas bessere statistische Eigenschaften aufweist als der ursprüngliche Stichprobenplan von Fritschi et al. (1976). Unab-

hängig davon, ob man jetzt das Originalverfahren verwendet oder das hier vorgeschlagene vereinfachte Verfahren, zeigt sich der Berner Stichprobenplan vor allem dann gegenüber der Klumpenstichprobe vorteilhaft, wenn ein Zusammenhang besteht zwischen dem Untersuchungsmerkmal und der Gemeindegröße. Bei Merkmalen ohne Zusammenhang zur Gemeindegröße sind die Verhältnisse etwas komplizierter. Die genaue Struktur der Intra-Klassen-Korrelation ist dann von Bedeutung. Etwas überraschend erzielt der Berner Stichprobenplan besonders gute Ergebnisse, wenn die interne Homogenität in den großen Gemeinden am stärksten ist. Dies erscheint etwas ungünstig, da man wohl eher davon ausgehen würde, dass, falls überhaupt Unterschiede bestehen, für die meisten Merkmale die interne Homogenität in kleinen, ländlichen Gemeinden stärker ist als in den Städten. Trotz dieses Vorbehalts zeigen die Simulationen jedoch, dass der Berner Stichprobenplan in den meisten Fällen deutlich besser abschneidet als ein einfaches Klumpungsverfahren und somit bei etwa gleich hohem Aufwand für die Beschaffung der Adressen der Zielpersonen insgesamt zu effizienteren Stichproben führt. Zudem ist nicht auszuschließen, dass sich die Stichprobeneffizienz noch weiter erhöhen lässt, wenn das Berner Stichprobenverfahren mit anderen Techniken wie zum Beispiel einer geschichteten Klumpenstichprobe auf Grundlage von Gemeindetypologien und -größen (vergleiche etwa Buchmann und Sacchi 1997) kombiniert wird. Dies wäre allenfalls in einer Folgestudie zu klären.

Auch gestaltet sich die Datenanalyse beim Berner Stichprobenplan kaum komplizierter als bei einer einfachen Klumpenstichprobe. Bei beiden Verfahren müssen Schätzer verwendet werden, die die Klumpenstruktur der Daten in Rechnung stellen, um die Varianzen der statistischen Koeffizienten nicht zu unterschätzen. Bei Stichproben auf Grundlage des *ex ante* geteilten Verfahrens ist zudem zu berücksichtigen, dass die Stichprobe aus zwei Schichten gezogen wurde, was die Varianzen im Allgemeinen verringert. Komplexe Schätzer, die Klumpungs- und Schichtungseffekte in angemessener Weise behandeln, stehen für eine Vielzahl statistischer Verfahren zur Verfügung (zur Analyse von komplexen Stichproben vgl. z.B. Lohr 1999 oder Levy und Lemeshow 2002) und sind mittlerweile in allgemeinen Statistik-Programmen recht gut unterstützt (vgl. z.B. StataCorp 2005). Surveys, bei denen es weder Schichtungs- oder Klumpenstrukturen noch Gewichte zur Korrektur unterschiedlicher Auswahlwahrscheinlichkeiten zu berücksichtigen gilt, sind ohnehin recht selten. Die Datenanalyse dürfte somit beim Berner Stichprobenplan keine zusätzlichen Schwierigkeiten bereiten.

Bei allen Vorteilen des Verfahrens ist eine zentrale Anwendungsvoraussetzung des Berner Stichprobenplanes jedoch noch zu nennen. Die hier präsentierten Simulationsergebnisse beruhen alle auf dem Gemeindeverzeichnis 2002 der Schweiz. Eine Eigenschaft der Schweizer Gemeinden ist, dass die Verteilung der Anzahl Personen, die in diesen Gemeinden leben, sehr schief ist. Das heißt, es gibt hunderte kleiner Gemeinden aber nur eine Hand voll große. Natürlich ist eine

solche Datenstruktur eine Bedingung dafür, dass der Berner Stichprobenplan überhaupt zu anderen Resultaten führt als ein einfaches Klumpungsverfahren. Wird das geteilte Stichprobenverfahren auf Aggregateinheiten angewendet, die sich in ihrer Größe nur wenig unterscheiden, ist die Wahrscheinlichkeit groß, dass keine einzige Einheit die notwendige Mindestzahl an Treffern für die Aufnahme in den ungeklumpten Zweig der Stichprobe erreicht. Eine schiefe Verteilung der Größen der verwendeten Aggregateinheiten ist also eine Voraussetzung für eine sinnvolle Anwendung des Berner Stichprobenplans. Dies lässt zum Beispiel den Einsatz des Verfahrens auf Ebene von Stimmbezirken, deren Größen in der Regel nicht so stark schwanken, als weniger günstig erscheinen.

5 Anhang

Die folgenden Ergebnisse werden zur Lösung von Gleichung (2) benötigt. Für eine binomialverteilte Zufallsvariable $X \sim B(n, p)$ ist die Wahrscheinlichkeit, k oder mehr Treffer zu erreichen, bekanntlich gegeben als

$$P(X \geq k) = \sum_{x=k}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$

(vgl. z.B. Evans et al. 2000, S. 43-47). Funktionen zur automatischen Berechnung von $P(X \geq k)$ stehen in den meisten Statistikprogrammen zur Verfügung. Etwas aufwändiger ist die Berechnung von $E(X|X \geq k)$, also dem Erwartungswert einer binomialverteilten Zufallsvariablen, gegeben eine bestimmte Mindestzahl an Treffern wird erreicht. Der Erwartungswert einer diskreten Variablen entspricht dem Mittel der mit den Auftretenswahrscheinlichkeiten gewichteten Ausprägungen, im vorliegenden Fall also

$$E(X|X \geq k) = \frac{k \cdot P(X = k) + (k+1) \cdot P(X = k+1) + \ldots + n \cdot P(X = n)}{P(X \geq k)}$$

was sich vereinfachen lässt zu

$$E(X|X \geq k) = k + \frac{P(X \geq k+1) + \ldots + P(X = n)}{P(X \geq k)}$$

6 Literaturverzeichnis

Babbie, E. R. (1979): The Practice of Social Research. Second Edition. Belmont, CA: Wadsworth.

Buchmann, M.; Sacchi, S. (1997): Berufsverlauf und Berufsidentität im soziotechnischen Wandel. Konzeption, Methodik und Repräsentativität einer retrospektiven Befragung der Geburtsjahrgänge 1949-51 und 1959-61. ETH Zürich.

Bundesamt für Statistik (2002): Gemeinde- und Ortschaftenverzeichnis 2002. Neuchâtel: BFS.

Evans, M.; Hastings, N.; Peacock, B. (2000): Statistical Distributions. Third Edition. New York: Wiley.

Fritschi, P.; Meyer, R.; Schweizer, W. (1976): Ein neuer Stichprobenplan für ein gesamtschweizerisches Sample. In: Schweizerische Zeitschrift für Soziologie 2(3), S. 149–158.

Gabler, S.; Häder, S. (1997): Überlegungen zu einem Stichprobendesign für Telefonumfragen in Deutschland. ZUMA-Nachrichten 41, S. 7-19.

Häder, S.; Glemser, A. (2006): Stichprobenziehung für Telefonumfragen in Deutschland. In: Diekmann, A. (Hrsg). Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Wiesbaden: VS Verlag, S. 148-171.

Handcock, M. S.; Morris, M. (1999): Relative Distribution Methods in the Social Sciences. New York: Springer.

Jann, B. (2001): Stichprobenziehung aus TwixTel. Universität Bern. Erhältlich unter http://www.socio.ethz.ch/people/jannb/wp/stichprobenziehung2up.pdf.

Kish, L. (1965): Survey Sampling. New York: Wiley.

Leu, R. E.; Burri, S.; Priester, T. (1997): Lebensqualität und Armut in der Schweiz. Bern: Haupt.

Levy, P. S.; Lemeshow, S. (1999): Sampling of Populations: Methods and Applications, 3rd ed. New York: Wiley.

Lohr, S.L. (1999): Sampling: Design and Analysis. Pacific Grove: Duxbury Press.

Meyer, R.; Haltiner, K.; Hofer, R.; Iff, H.; Rüegg, W. (1982): Fragen an die Zukunft. Die Bedeutung von Beruf, Bildung und Politik für die zwanzigjährigen Schweizerinnen und Schweizer. Aarau und Frankfurt am Main: Sauerländer.

Sasieni, P. (1998): An adaptive variable span running Line smoother. In: Stata Technical Bulletin 41, S. 4-7.

Schmugge, S.; Grau, P. (1998): Telefon-Anschlüsse der privaten Haushalte in der Schweiz. Situationsanalyse 1998. Luzern: LINK Institut.

Schmugge, S.; Grau, P. (2000): Sind die SchweizerInnen überhaupt noch zu erreichen? Telefonanschlüsse der privaten Haushalte in der Schweiz im Jahr 2000. Luzern: LINK Institut.

Schnell, R. (1991): Wer ist das Volk? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen“: Undercoverage, Schwererreichbare und Nichtbefragbare. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 43 (1), S. 106-137.

Schnell, R. (1997): Praktische Ziehung von Zufallsstichproben für Telefon-Surveys. In: ZA-Informationen 40, S. 45-59.

StataCorp (2005): Stata Survey Data Reference Manual. College Station, Texas: Stata Press.

Wydler, H.; Walter, T.; Hattich, A.; Hornung, R.; Gutzwiller, F. (1996): Die Gesundheit 20jähriger in der Schweiz: Ergebnisse der PRP 1993. Aarau: Sauerländer.

Der Ipsos SOWI-Bus: Stichprobenanlage und erste Untersuchungsergebnisse

Christian Holst

Für die Akzeptanz von Befragungsergebnissen in der wissenschaftlichen wie allgemeinen Öffentlichkeit ist der Nachweis von Qualität entscheidend. Über die Definition dessen, was Qualität in Bezug auf sozialwissenschaftliche Erhebungen jedoch ist, wird noch trefflich gestritten. Allein über die Kriterien „richtige Ergebnisse“ und „korrekte Verfahren“ wird ein Qualitätsnachweis nicht erfolgreich sein, da die Profession mittlerweile eine ganze Reihe korrekter Verfahren zur Verfügung stellt. An dieser Stelle wird statt dessen argumentiert, dass die Qualität einer Studie danach zu bemessen ist, ob die verwendeten Verfahren adäquat sind, ein bestimmtes Erkenntnisinteresse zu verfolgen. Über die unterschiedlichen Erkenntnisinteressen ergeben sich dann jeweils unterschiedliche Anforderungen, die sich anhand der Dimensionen angestrebte Genauigkeit, verfügbare Zeit und verfügbares Budget einordnen lassen. Die gängigsten zur Verfügung stehenden Stichprobendesigns werden anschließend kurz beschrieben und in Bezug auf ihre Qualitätsmerkmale eingeordnet. In einem zweiten Teil wird der Ipsos SOWI-Bus als sozialwissenschaftliche Mehrthemenumfrage vorgestellt, die qualitativ im oberen Feld der Random-Walk-Verfahren angesiedelt ist. In einem Vergleich einiger zentraler Variablen mit dem ALLBUS wird die Qualität des Ipsos SOWI-Bus überprüft.

1 Einleitung

Die Qualität sozialwissenschaftlicher Umfragen ist – seit einiger Zeit – wieder zum Thema geworden. Die jetzige gemeinsame Tagung von ASI und der Methodensektion der DGS ist sicherlich auch nur ein weiterer Schritt zum Bemühen um Exzellenz in der Erhebungsqualität, die Polemik von Mohler, Koch und Gabler (2003), die Analyse von Schneekloth und Leven (2003), und nicht zuletzt die Denkschrift der DFG (1999) sowie die gemeinsamen Standards zur Qualitätssicherung in der Markt- und Sozialforschung von ADM, ASI und BVM (1999) sind Ausdruck dieser anhaltenden Diskussion. Darüber hinaus wurden mit der Verleihung des Europäischen René-Descartes-Preises an das European Social Survey auch die Bemühungen um eine höchstmögliche Datenqualität in vergleichenden europäischen sozialwissenschaftlichen Erhebungen gewürdigt und ge-

zeigt, dass das Streben nach bestmöglicher Datenqualität nicht nur innerhalb der Profession, sondern auch außerhalb anerkannt wird.

Die Anerkennung wissenschaftlicher Ergebnisse sowohl innerhalb der Profession wie auch außerhalb ist prinzipiell voraussetzungsvoll. Da letztlich das Zustandekommen einzelner Ergebnisse für Außenstehende nicht oder nur kaum nachvollziehbar ist, ist umgekehrt die Versicherung einer höchstmöglichen Qualität von Stichprobenziehung, Durchführung und Auswertung das entscheidende Argument für die Anerkennung der Ergebnisse bei Auftraggebern und Öffentlichkeit. Diese Versicherung steckt aber in einem Dilemma: selbstverständlich wird erwartet, dass die Qualitätsanforderungen „höchstmöglich" sind, die entscheidende Frage ist jedoch erstens, welches die Qualitätsanforderungen sind, und zweitens, was unter gegebenen Umständen möglich ist.

Damit gewinnt die Methodologie als Regelwerk und Referenzmaßstab nicht nur innerhalb der Wissenschaft, sondern auch außerhalb an Bedeutung. So gilt bislang die Ausschöpfungsquote als ein „Kernindikator" für die Güte von Erhebungen, weil sie erstens davon ausgeht, dass nur durch eine hohe Ausschöpfung Verzerrungen in der Stichprobe minimiert sind und somit die Stichprobe auch eine wirklichkeitsgetreue Abbildung der Grundgesamtheit darstellt (Repräsentativität), zweitens aber auch durch ihre Reduzierung auf einen Prozentwert eine einfache Vergleichbarkeit verschiedener Untersuchungen impliziert wird, nach der dann die eine Untersuchung „besser" oder „schlechter" ist als die andere. In der sozialwissenschaftlichen Diskussion um Sampling- und Non-Sampling-Fehler ist dabei längst deutlich geworden, dass eine Reduzierung der Qualitätsindikatoren auf nur eine Kennziffer an der Wirklichkeit vorbei geht (vgl. DFG 1999: 94), und dass andere, methodologische Aspekte eine mindestens ebenso wichtige, wenn nicht entscheidendere Rolle spielen.

2 Aspekte von Qualität in Bevölkerungsumfragen

Im Rahmen der DFG-Denkschrift über „Qualitätskriterien in der Umfrageforschung" (1999) wurden als absolute und immer gültige Kriterien für die Güte einer Umfrage gefordert

- „(…) die verzerrungsfreie Abbildung einer definierten Grundgesamtheit durch die Stichprobe;
- die gültige, zuverlässige Messung der gemeinten Sachverhalte durch die Befragung" (a.a.O., S. 94).

Daraus ergeben sich zwei instrumentelle Kriterien, die die Qualität von Umfragen beschreiben sollen:

- „das erste sind richtige Ergebnisse;

- das zweite sind korrekte Verfahren“ (a.a.O.).

Beides ist selbstverständlich richtig – niemand käme auf die Idee, viel Geld für empirische Studien auszugeben, wenn er nicht „richtige“ Ergebnisse erwarten würde. Ebenso ist die Anwendung korrekter Verfahren überhaupt die wesentliche Voraussetzung dafür, richtige Ergebnisse zu erhalten. Denn nur durch korrekte Verfahren können die erzielten Ergebnisse überhaupt den Anspruch erhalten, als „richtig“ zu gelten, da ihre Verteilungen ja in der Regel bislang unbekannt waren. Insofern ist die Forderung nach der ausführlichen Dokumentation der eingesetzten Verfahren ebenso alt – sie wurde bereits 1903 auf der Tagung des „International Statistics Institute“ erhoben (vgl. Quatember 2001: S. 6) – wie aktuell.

Diese von der DFG erhobene Qualitätsdefinition ist jedoch in gewisser Weise unbefriedigend, denn sie liefert keinen Maßstab dafür, ab wann und unter welchen Bedingungen von Qualität gesprochen werden kann. Zunächst besteht ein enger Zusammenhang zwischen dem Erreichen eines „richtigen“ Ergebnisses und den eingesetzten Verfahren. Neue, bislang unbekannte Erkenntnisse können nur insoweit Gültigkeit und Akzeptanz in der Öffentlichkeit beanspruchen, wie ihr Zustandekommen dokumentiert und nachprüfbar ist – seien es Resultate sozialwissenschaftlicher Befragungen oder naturwissenschaftlicher Laborversuche. Durch die Akzeptanz von als korrekt angesehenen Verfahren werden auch die erzielten Ergebnisse gewissermaßen „geadelt“ und akzeptanzfähig. Nicht zuletzt ist es in der Diskussion um (unbequeme) Befragungsergebnisse eine gängige Strategie, zunächst einmal die eingesetzten Verfahren in Frage zu stellen (vgl. aktuell Daves/ Newport (2005)).

Das verweist jedoch wiederum auf die Bedeutung dessen, was als „korrektes“ Verfahren angesehen wird. Die Methodologie besonders in der Umfrageforschung entwickelt sich – wie jeder andere Wissenschaftszweig – ständig weiter, sei es von Außen durch neue technische Möglichkeiten (z.B. Telefon, Laptop, Online), oder von Innen durch Entwicklungen z.B. in Stichprobenverfahren (die weitgehende Ablösung von Quoten- durch Randomstichproben bei allgemeinen Bevölkerungsbefragungen), Erhebungsdesigns, Skalen etc. Korrekte Verfahren werden modifiziert, weiterentwickelt und/oder durch neue abgelöst. Durch diese Entwicklungen ist eine Vielfalt verschiedener Verfahren entstanden, die alle durchaus korrekt sein können. Ein „Methodendogmatismus“, der nur das eine Verfahren bestehen lässt, kann sich angesichts dieser Pluralisierung im verfügbaren Methodenkanon nicht mehr halten.

Somit ist stattdessen immer zu fragen, welches Verfahren unter den verfügbaren das geeignete ist, um ein bestimmtes Problem zu beschreiben. Die Wahl des korrekten Verfahrens richtet sich damit nach den Forschungsabsichten bzw. dem Erkenntnisinteresse. Diese können jedoch sehr unterschiedlich ausfallen: Sie reichen von Informationen über Verteilungen, die sehr schnell verfügbar sein müssen (um auf eine gemessene Erwartungshaltung z.B. schnell und adäquat reagie-

ren zu können), bis zu sozialwissenschaftlichen Untersuchungen, bei denen komplexe soziale Phänomene erforscht werden, deren Häufigkeit und Struktur nur theoretisch vermutet, aber empirisch noch nicht nachgewiesen wurden. Steht bei den einen die Zeit als ausschlaggebender Parameter im Vordergrund, ist es bei den anderen die Genauigkeit der Messung. Die Ergebnisse dieser verschiedenen Studien sind damit immer auch im Kontext ihrer Entstehung und der Forschungsabsicht zu interpretieren: So wäre es gleichermaßen falsch, an eine unter Zeitdruck durchgeführten Untersuchung denselben Anspruch an Präzision zu stellen, wie an eine unter der Prämisse der Genauigkeit durchgeführten Untersuchung, dass sie sehr schnell erste, handlungsleitende Ergebnisse liefern kann. Die Spanne der Erhebungszeiten reicht mittlerweile von in einer Nacht durchgeführten Befragungen, bei denen die Ergebnisse am nächsten Tag zur Verfügung stehen, bis zu sozialwissenschaftlichen Studien wie dem ALLBUS oder dem SOEP, bei denen die Feldarbeit bis elf Monate dauern kann.

Damit ergeben sich je nach Untersuchungsabsicht auch unterschiedliche Qualitätsanforderungen, die unter der Prämisse „Fitness for use“ (Biemer/Lyberg 2003: 13) subsumiert werden können. Die Qualität einer Umfrage lässt sich danach in drei Dimensionen beschreiben. Sie muss

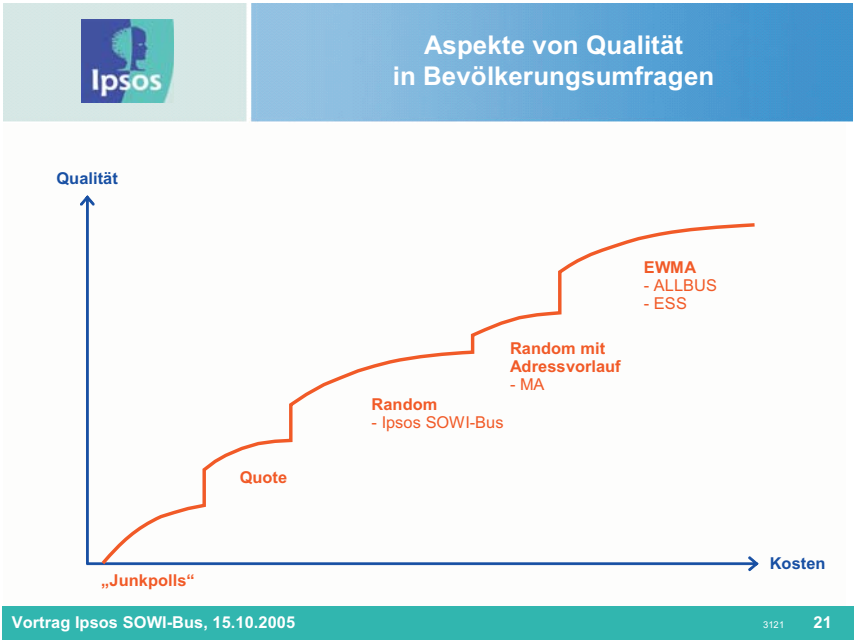
- so *genau* wie nötig sein, um den angestrebten Zweck zu erfüllen;
- *pünktlich* zu dem Zeitpunkt fertig gestellt sein, zu dem sie benötigt wird;
- und *zugänglich* für diejenigen sein, für die sie erstellt wurde.

Während sich der dritte Punkt eher an den Bedürfnissen der amtlichen Statistik ausrichtet und weder für die akademische noch die betriebliche Umfragepraxis relevant ist (da hier immer auch ein Auftraggeber-/Auftragnehmer-Verhältnis existiert und der Auftraggeber auch die beauftragte Umfrage erhält), beschreiben die ersten beiden genannten Punkte der Genauigkeit und Pünktlichkeit zwei der drei Entstehungsparameter, unter denen die Umfragepraxis heute operiert.

Der dritte, bislang noch nicht angesprochene Parameter, ist letztlich das für eine Untersuchung zur Verfügung stehende Forschungsbudget. Das Budget bemisst nicht nur die Zahl oder die Länge der zu erzielenden Interviews, sondern auch die vielfältigen Kontroll- und Qualitätssicherungsschritte, die ein Institut zur Herstellung von Genauigkeit und Pünktlichkeit unternehmen kann. Dies beginnt bei den Kosten der Stichprobenbildung (Einwohnermeldeamtsstichproben sind allein durch die Gebührenordnungen der Städte und Gemeinden per se teurer als Zufallsstichproben nach dem ADM-Verfahren), geht über die Kosten der Interviewerschulung (persönliche Interviewerschulungen von 100 und mehr Interviewern bei Face-to-face-Befragungen sind allein durch die Reisekosten und Tagegelder teurer als eine schriftliche Schulung), Interviewerhonorare und -nebenkosten in Abhängigkeit z.B. von zu erbringenden Kontaktversuchen, ausschöpfungsverbessernden Maßnahmen wie z.B. einer telefonischen Re-Kontaktierung

von zunächst nicht teilnahmebereiten Befragten, bis zu nachträglichen Kontaktaufnahmen mit Befragten zur Klärung von nicht plausiblen oder inkonsistenten Befragungsdaten. Die Unterschiede in den Fallpreisen der ALLBUS-Erhebungen zwischen 1980 von 75 DM und 250 DM im Jahr 2000 reflektieren nicht nur eine allgemeine Preissteigerung, sondern eben auch Unterschiede in der Untersuchungsanlage (vgl. Koch 2002: Abb. 3). Wahrscheinlich nur in seltenen Fällen wird es möglich sein, ein optimales Forschungsdesign ohne Rücksicht auf Kosten zu beantragen und bewilligt zu bekommen - glücklicherweise sind aber nach wie vor Projekte wie der ALLBUS oder das European Social Survey möglich. Der größte Teil der Projekte wird aber bereits in der Konzeptionsphase abwägen müssen, welche Qualitätsmerkmale tatsächlich absolut notwendig sind („need to have"), und auf welche Qualitätsmerkmale möglicherweise verzichtet werden kann („nice to have"). Eine Rolle dabei dürften immanente Schranken und Regelungen in den Antragsverfahren der jeweiligen Stiftungen und Geldgeber sein, aber auch die Konkurrenz mit anderen Forschungsprojekten um endliche und damit knappe Fördermittel.

Diese Trias aus angestrebter Genauigkeit – verfügbarer Zeit – verfügbarem Budget definiert letztlich den möglichen Qualitätsrahmen, der in einem Forschungsprojekt ausgefüllt werden kann, und der den Kontext darstellt, innerhalb dessen die Ergebnisse interpretiert werden müssen. Eine der Aufgaben der Stichprobentheorie ist es schließlich zu versuchen, solche Auswahl- und Schätzverfahren zu entwickeln, die bei möglichst niedrigen Kosten Schätzungen liefern, die für das jeweilige Ziel genau genug sind (Cochran 1972: 24). Trägt man innerhalb dieses Rahmens aus Kosten und Genauigkeit die gängigen Stichprobenverfahren auf, ergibt sich idealtypisch ein asymptotischer Verlauf, der durch Schwellenwerte unterbrochen wird.



2.1 Willkürliche Auswahlen

Die unter Kostengesichtspunkten sicherlich billigsten, aber hinsichtlich der Genauigkeit auch unbefriedigendsten Stichproben sind dabei willkürliche Auswahlverfahren, wie es gemeinhin „Fußgängerzonenumfragen“, TED-Umfragen oder im Online-Bereich frei zugängliche „Online-Befragungen“ sind. Die Auswahl der Zielpersonen erfolgt oftmals selbstselektiv, d.h. es liegt in der Entscheidung der Befragten, ob sie teilnehmen oder nicht, oder, wenn Interviewer eingesetzt werden, folgt die Auswahl keinem festgelegten Algorithmus. Weder besteht eine definierte Grundgesamtheit noch ein nur annähernd beschreibbarer Stichprobenrahmen. Es werden zwar Daten gesammelt – die oft durch ihre hohe Fallzahl beeindrucken –, doch ein Rückschluss auf eine Grundgesamtheit ist nicht möglich. Die Kosten werden durch den oftmals fehlenden, oder, wenn vorhanden, unqualifizierten und unkontrollierten Einsatz von Befragern extrem niedrig gehalten. Die Dateneingabe erfolgt z.B. bei TED- oder Online-Umfragen maschinell, eine Überprüfung oder Plausibilitätskontrolle sowie Datenaufbereitung dürfte entweder gar nicht oder nur rudimentär existieren. Diese Art von Stichproben unterscheiden sich in ihren Ergebnissen signifikant von unter kontrollierten Bedingun-

gen erhobenen Daten (vgl. z.B. Faas 2003), sie sind im besten Falle nur „informativ“ zu nennen (Quatember 2001: 20), und genügen elementaren Regeln wissenschaftlicher Arbeit nicht (Schnell/Hill/Esser 1999: 278).

2.2 Quotenstichproben

Quotenstichproben stellen den oben beschriebenen Stichproben gegenüber einen deutlichen qualitativen Sprung in der Genauigkeit der Durchführung, aber auch den damit verbundenen Kosten dar. Während sie zwar im Bereich der allgemeinen Bevölkerungsumfragen weitgehend von Zufallsstichproben abgelöst wurden, spielen sie nach wie vor eine wichtige Rolle bei Untersuchungen von kleinen Zielgruppen. Per Definition lassen sich dabei für die als Quotenmerkmale erhobenen Variablen hohe Genauigkeiten erzielen. Während bei der Vorgabe von Randquoten diese Genauigkeiten nur in den Randverteilungen erzielt werden, lassen sich durch kombinierte oder verbundene Quoten diese Genauigkeiten auch in den Zellen herstellen. Je nach Anzahl, Definition und ggf. Kombination mehrerer Merkmale lassen sich innerhalb einer Stichprobe mehr oder weniger viele Zellen „genau“, d.h. entsprechend den als bekannt vorauszusetzenden Verteilungen in den Grundgesamtheiten, abbilden. Implizit wird dabei davon ausgegangen, dass bei einer hohen Anzahl „genau“ abgebildeter Quotenzellen auch die Verteilungen in den nicht-quotierten Zellen der Stichprobe den Verteilungen in der Grundgesamtheit entsprechen. Mit zunehmender Anzahl und Kombination der eingesetzten Quotenmerkmale sollte sich dann auch die Qualität der Stichprobe erhöhen. Eine solche Qualitätssteigerung führt allerdings auch zu höheren Kosten: Je mehr und je enger die Quotenvorgaben für die Interviewer gesetzt werden, desto schwieriger wird es, gegen Ende der Feldzeit Zielpersonen zu finden, die genau einer bestimmten Kombination von Merkmalen entsprechen. So ist die Erfüllung einfacher Randquoten relativ problemlos und damit vergleichsweise günstig, wohingegen komplexe Kombinationen z.B. aus Ortsgröße, Alter, Geschlecht und Beruf in der Bearbeitung aufwändig sind. Da die Interviewerhonorare immer auch den Aufwand berücksichtigen müssen, den die Interviewer für die Identifikation einer Zielperson haben, ziehen somit schwieriger zu erfüllende Quotenvorgaben auch höhere Kosten nach sich. Dabei ist davon auszugehen, dass das Verhältnis von Kosten und Qualitätssteigerung (z.B. im Sinne von zusätzlichen Quotenvorgaben) kaum linear, sondern sich eher asymptotisch einem bestimmten Niveau annähert: Der Übergang von Randquoten zu kombinierten Quoten stellt einen deutlich höheren Qualitätszuwachs dar als z.B. die Hinzunahme eines weiteren Merkmals bei kombinierten Quoten. So können auch bei einer strikten Einhaltung der Randquoten viele verschiedene zusammengesetzte Stichproben diese Bedingung erfüllen, während dies bei einer kombinierten Quotenvorgabe jeweils nur eine Stichprobenzusammensetzung kann (Schnell/Hill/Esser

1999: 281). Umgekehrt lässt jedoch eine beliebige Addition von weiteren kombinierten Quotenmerkmalen die Kosten exponentiell steigen, ohne dass ein damit einhergehender (oder proportionaler) Qualitätszuwachs zu erwarten wäre. Eine solche Quotenstichprobe würde dabei praktisch auf das Problem stoßen, dass die Suche nach den letzten Zielpersonen, die diese eine Kombination erfüllen, einer Suche nach der Nadel im Heuhaufen gleich käme.

Ebenso bemisst sich die Qualität einer Quotenstichprobe nicht zuletzt an der Qualität der verwendeten Quotenmerkmale: Die Quotenmerkmale müssen erstens für den Untersuchungszweck geeignet sein (d.h. es muss ein theoretisch angebbarer Zusammenhang zwischen Untersuchungsziel und Quotenmerkmalen hergestellt werden), sie müssen zweitens verfügbar sein (die Verteilungen in der Grundgesamtheit müssen überhaupt erst schon einmal gemessen worden sein), drittens aktuell sein (ansonsten würde eine vergangene Verteilung reproduziert werden), und viertens beobachtbar sein (d.h. die Interviewer müssen eine Chance haben, die Zielperson durch wenige gezielte Fragen in eine der Quotenzellen zuzuordnen). Wo dies nicht gegeben ist, werden auch korrekt erfüllte Quotenzellen keine Stichprobe bilden, die eine adäquate Beschreibung einer Grundgesamtheit zulässt. Schließlich würde eine Quotenstichprobe auch die theoretische „Schallgrenze“ nicht durchbrechen können, dass alle Methoden der schließenden Statistik – d.h. des Rückschlusses von einer Stichprobe auf die Grundgesamtheit – nur für Zufallsstichproben definiert sind. So lässt sich nur durch Zufallsstichproben auch die *Genauigkeit* angeben, mit der aus einer Stichprobe auf eine Grundgesamtheit geschlossen werden kann. Insofern stoßen auch sehr aufwändige Quotenstichproben an immanente Grenzen, die auch bei hohen Kosten nur ein bestimmtes Qualitätsniveau erlauben.

2.3 Zufallsstichproben: Random Route

Damit schließen unter Qualitätsaspekten Zufallsstichproben als weiterer Schritt an Quotenstichproben an, weil allein für sie der Repräsentationsschluss von einer Stichprobe auf eine Grundgesamtheit definiert und die Genauigkeit von Schätzern berechenbar ist. Zufallsstichproben sind mittlerweile unter Bevölkerungsstichproben in der Bundesrepublik der etablierte Standard. Mit dem ADM-Stichprobensystem für Face-to-face-Stichproben sowie dem auf dem Gabler-Häder-Design beruhenden ADM-Stichprobensystem für Telefonstichproben stehen der Sozial- und Marktforschung zwei sehr leistungsfähige und qualitativ hochwertige Stichprobenverfahren mit den dazugehörigen Ziehungsprogrammen für Zufallsstichproben nach Random-Walk- bzw. Random-Dial-Verfahren zur Verfügung (vgl. ADM 1979; ADM/AG.MA 1999; Althoff 1993; Gabler / Häder / Hoffmeyer-Zlotnik 1998). Die Genauigkeit dieser Stichprobenverfahren lässt sich auch hier wieder vom „klassischen“ Random-Walk-Verfahren bis zu komplexen

Einwohnermeldeamtsstichproben gruppieren, wobei auch hier wieder qualitative Sprünge in der Verfahrensweise einzelner Systeme zu konstatieren sind. Im Wesentlichen spielt dabei eine Rolle, welchen Spielraum die Interviewer auf den letzten beiden Stufen des Auswahlprozesses, also der Haushalts- wie der Zielpersonenauswahl, haben können. Während im „klassischen" Random-Route den Interviewern vom Institut der Sample-Point als Adresse vorgegeben wird, die Haushalts- und die Zielpersonenauswahl aber entsprechend den Institutsanweisungen durch den Interviewer in einem Gang vorgenommen wird, erfolgt im „Random-Route mit Adressvorlauf" die Haushaltsauswahl in einem separaten Schritt. Dabei wird dem Institut eine Auflistung der Adressen, die sich entsprechend der Random-Route ergeben, wieder zurückgesendet, aus der dann das Institut bestimmte, zu kontaktierende Adressen auswählt. Damit erstellt das Institut eine Adressstichprobe, aus der dann dem Interviewer, der das Interview durchführen soll, konkrete Adressen vorgegeben werden. Die korrekte Haushaltsauswahl ist somit durch das Institut weitgehend kontrollierbar geworden, darüber hinaus ist somit ein Stichprobenrahmen entstanden, aus dem die Ausschöpfung der Adressen auch zuverlässig berechnet werden kann. Der doppelte Interviewereinsatz führt einerseits zu höheren Kosten als der effizientere Einsatz im klassischen Random-Route, erlaubt andererseits aber auch eine höhere Kontrolle und Nachvollziehbarkeit der Selektionsmechanismen auf der zweiten Auswahlstufe.

2.4 Zufallsstichproben: Einwohnermeldeamtsstichproben

Eine weitere Steigerung der Stichprobenqualität wird durch die Verwendung von Einwohnermeldeamtsstichproben gegeben. Hier wird der Spielraum der Interviewer völlig eingeschränkt und nachvollziehbar. Durch die Ziehung von Personen aus den Melderegistern entfallen sowohl die Haushalts- wie die Zielpersonenauswahl, letztere wird den Interviewern namentlich genannt, so dass ein Ausweichen auf andere Personen auch nicht möglich ist. Damit ist der komplette Prozess der Stichprobenbildung vollständig nachvollziehbar und beschreibbar. Studien, deren Stichprobe auf dieser Grundlage gebildet wurden, entsprechen am ehesten der „best practice" sozialwissenschaftlicher Forschung. Diese hat allerdings auch ihren Preis, denn nicht nur ist in den Instituten ein erheblicher Mehraufwand an Interviewerschulung, -kontrolle und –koordination zu leisten, sondern auch die Feldvorbereitung – z.B. die Koordination mit Meldebehörden, Abgleich und Kontrolle der erhaltenen Adressstichproben – sowie die von den Behörden erhobenen Gebühren für die Stichprobenziehung schlagen dabei zu Buche. Werden dann – wie auch in herkömmlichen ADM-Stichprobenverfahren möglich – noch zusätzliche ausschöpfungsverbessernde Maßnahmen wie die Konversion von zunächst nicht befragungsbereiten Personen (vgl. dazu Neller 2005), die Nachrecherche von verzogenen Personen etc. durchgeführt, können die Kosten extrem

steigen. Nicht zuletzt wird deswegen auch von Seiten der Auftraggeber gefragt, inwieweit diese Kosten durch die Verbesserung der Datenqualität gegenüber herkömmlichen ADM-Stichproben noch gerechtfertigt sind (Koch 2002: 51; Gabriel/Keil 2005: 616).

Für Bevölkerungsumfragen stehen somit eine ganze Reihe von unterschiedlichen Stichprobenverfahren mit jeweils unterschiedlicher Güte zur Verfügung. Wesentlich für die Beurteilung ihrer Güte ist dabei die Transparenz des Ziehungs- und Erhebungsvorganges. Auf der Grundlage dieser Informationen kann dann eine qualifizierte Entscheidung getroffen werden, welche Güte der Daten für den jeweiligen Untersuchungszweck benötigt wird, und welche Kosten man dafür bereit ist, zu bezahlen.

Mit dem Ipsos SOWI-Bus hat Ipsos deshalb ein Instrument wieder aufgelegt, das den Sozialwissenschaften auf der einen Seite eine qualitativ hochwertige Zufallsstichprobe nach dem ADM-Verfahren bietet, auf der anderen Seite aber die Kostenseite im Blick behält. Eine erste Welle ist im Sommer 2005 im Feld gewesen, im Folgenden soll über die Anlage des SOWI-Bus und die Datengüte berichtet werden.

3 Der Ipsos SOWI-BUS

3.1 Vorgänger, Hintergrund und Zielsetzung

Der Ipsos SOWI-Bus ist eine speziell auf die Bedürfnisse der Sozialwissenschaften ausgerichtete halbjährliche Umfrage, die auch die Einschaltung von nur kurzen Fragenblöcken mit sozialwissenschaftlichem Inhalt ermöglicht. Sie knüpft in der Anlage an den in den neunziger Jahren von ZUMA verantworteten und von unserem Vorläuferinstitut GFM/Getas WBA durchgeführten SOWI-Bus an (vgl. v. Harder / Hoffmeyer-Zlotnik 1990, Hoffmeyer-Zlotknik 1997). Dieser Bus ist – schon der Name deutet dies an – eine Omnibus- oder Mehrthemenumfrage, d.h. es beteiligen sich an einer Erhebung mehrere Auftraggeber, die auch jeweils unterschiedliche Themen in diese Umfrage einbringen. Die Anlage dieser Studie wird von Ipsos Public Affairs / Politik- und Sozialforschung entworfen und verantwortet, die Einflussmöglichkeiten der Auftraggeber sind dabei – anders als bei ad hoc-Umfragen – nur begrenzt. Dies gilt insbesondere für Art und Termin der Durchführung sowie die erhobene Soziodemografie. Für diese Einschränkungen werden im Gegenzug die Möglichkeiten zur Erhebung eigener Fragenprogramme gegeben zu einem Budget, das deutlich unter dem einer eigenen ad hoc-Erhebung mit vergleichbarem Qualitätsstandard bleibt.

Während Omnibusumfragen in der Marktforschung ein gängiges und bewährtes Instrument sind, sind sie im Bereich der Sozialwissenschaften eher unbedeu-

tend und mit Vorurteilen belastet. Die gängigsten sind wohl die Einwände, dass man selber keine Kontrolle auf das Umfeld habe, in dem seine Fragen gestellt werden, und somit unliebsame und nicht kontrollierbare Kontexteffekte gewärtigen müsste, sowie dass die Qualität der Stichproben von Omnibusumfragen nicht den Ansprüchen sozialwissenschaftlicher Forschung genügten. Auf beide Vorurteile wird im Ipsos SOWI-Bus ausdrücklich eingegangen: So erhalten die Beteiligten eine Übersicht der in der laufenden Welle erhobenen Themenblöcke und können auf diese Weise feststellen, in welchem Kontext ihre eigenen Fragen standen. Mit der Beschränkung auf ausschließlich sozialwissenschaftliche Themen entfällt auch die Befürchtung, dass die eigenen Fragen völlig themenfremd „zwischen Windeln und Müsli" erhoben würden. Hinsichtlich der Stichprobenqualität ist es eines der kennzeichnenden Merkmale des Ipsos SOWI-Bus, dass Stichprobenziehung, Feldarbeit und Datenaufbereitung wie bei anderen anspruchsvollen sozialwissenschaftlichen ad hoc-Studien in sämtlichen Schritten dokumentiert und transparent gemacht werden. In seiner Anlage und Durchführung wird besonderer Wert auf hohe methodische Qualität gelegt, und diese wird auch dokumentiert.

Der Ipsos SOWI-Bus ist dafür konzipiert worden, bei begrenzten Fragenprogrammen, für die eine eigene ad hoc-Erhebung zu aufwändig wäre, ein einer ad hoc-Studie vergleichbares Ergebnis zu liefern. Gerade unter dem Aspekt auch schrumpfender Forschungsbudgets schließt er damit eine Lücke zwischen kleinen, oftmals in Eigenregie und bei einer räumlich, qualitativ und methodisch eingeschränkten Stichprobe durchgeführten Erhebungen einerseits, sowie umfangreichen und teuren ad hoc Befragungen. Insofern ist er ideal geeignet beispielsweise für Inzidenzmessungen im Vorfeld größerer Untersuchungen, als Pretest für neu entwickelte Erhebungsinstrumente, für Skalentests oder für kontinuierliche Erhebungen (Trackings) zur Beobachtung von Veränderungen in zentralen Variablen.

3.2 Untersuchungsdesign und Methode

Der Ipsos SOWI-Bus wird als persönliche, computergestützte Face-to-face-Befragung (CAPI) durchgeführt. Insgesamt werden ca. 1.000 Interviews erhoben, davon ca. 800 in den alten Bundesländern (einschließlich West-Berlins), und ca. 200 in den neuen Bundesländern (einschließlich Ost-Berlins)[1]. Prinzipiell ist eine Aufstockung der Fallzahlen auf 1.500 oder als „Doppelwelle" mit 2.000 Interviews möglich. Die Grundgesamtheit des Ipsos SOWI-Busses ist die deutschsprachige, in Privathaushalten lebende Wohnbevölkerung ab 18 Jahren der Bun-

1 Durch die Neueinteilung der Bezirke in Berlin, die in einigen Bezirken nunmehr Gebiete sowohl des West- wie des Ostteils der Stadt umfassen, ist bei einigen wenigen Befragten eine eindeutige Zuordnung zur Ost- bzw. Weststichprobe nicht mehr möglich.

desrepublik Deutschland. Um den wissenschaftlichen Ansprüchen einer Zufallsstichprobe zu genügen, wird ein mehrstufig geschichtetes Auswahlverfahren auf der Grundlage des ADM-Stichprobensystems 2003 mit limitierter Adressauswahl ohne Vorabbegehung (Random-Route) verwendet. Daraus wird ein Netz, d.h. 258 Sample-Points, gezogen, wobei entsprechend 210 Sample-Points auf den Westen und 48 Sample-Points auf den Osten entfallen. Die Stichprobe wird nach Bundesländern und BIK-Ortsgröße geschichtet. Bei diesem Bruttoansatz sind somit im Mittel etwa 3,8 Netto-Interviews pro Point im Westen und 4,2 Netto-Interviews pro Point im Osten zu erzielen. Selbstverständlich wird für jede Erhebungswelle jeweils eine neue Stichprobe gezogen. Wenn in den ausgewählten Sample-Points trotz intensiver Bemühungen kein Erfolg zu erzielen ist oder der Sample-Point so abseitig liegt, dass er nur unter extremem Aufwand zu erreichen ist, können insgesamt bis zu fünf Prozent der Sample-Points strukturneutral (d.h. innerhalb des gleichen Regierungsbezirks und der gleichen BIK-Ortsgröße) ersetzt werden. Die Ersetzungen werden vom Institut dokumentiert.

Die Auswahl der Zielhaushalte innerhalb der einzelnen Sample-Points erfolgt durch das Random-Route-Verfahren mit limitierter Adressauswahl (8 aus 23) ohne Vorabbegehung. Die Startadressen der jeweiligen Sample-Points werden im Institut per Zufallsauswahl aus der Startadressendatei ermittelt und auf das Kontaktprotokoll übertragen. Ausgehend von dieser zufällig ermittelten Startadresse listet der Interviewer nach einer eindeutig festgelegten Begehungsvorschrift (Random-Route-Verfahren) 23 Privathaushalte auf. Aus diesen Haushalten sind acht, vom Institut bereits vorab markierte, Zielhaushalte in die Befragung einzubeziehen.

Die Interviewer führen ein Protokoll, in dem Zahl, Art und Erfolg der Kontakte vermerkt werden. Können einzelne Interviewer nicht die erforderliche Zahl von durchschnittlich fünf Interviews je Sample-Point erfüllen, kann die Feldleitung entscheiden, ob noch zusätzliche Adressen von der Adressliste eingesetzt werden dürfen. Im Durchschnitt der Sample-Points dürfen maximal zwei Zusatzadressen aufgenommen werden. In der Feldleitung wird dokumentiert, in welchen Fällen und warum Zusatzadressen aufgenommen werden.

Hat der Interviewer den Zielhaushalt erreicht, so werden auf der dritten Auswahlstufe in denjenigen Haushalten, in denen mehr als eine potenzielle Zielperson (deutschsprechende Wohnbevölkerung im Alter ab 18 Jahren) lebt, die Haushaltsmitglieder aufgelistet und die eigentliche Befragungsperson durch einen Zufallszahlenschlüssel bestimmt. Mit dieser Person wird das Interview durchgeführt werden. Die Interviewer werden innerhalb von drei Kontaktversuchen zu unterschiedlichen Zeiten und an unterschiedlichen Tagen versuchen, einen persönlichen Kontakt mit der Zielperson herzustellen und das Interview durchzuführen.

Je Erhebungswelle werden ca. 250 Interviewer eingesetzt. Dies sind Interviewer, die sich in der Vergangenheit bei der Durchführung von Face-to-face-Umfragen besonders qualifiziert haben. 30% der Nettointerviews werden durch die Feldkontrolle überprüft. Die Kontrollen werden so angelegt, dass jeder der Interviewer mindestens einmal kontrolliert wird. Die Kontrollen der Netto-Interviews finden postalisch mit den Zielpersonen statt und beinhalten Fragen zum Thema und dem Zeitpunkt des Interviewerbesuches. Kommt ein solcher postalischer Kontakt nicht zustande, so werden die Kontrollen – wo möglich – telefonisch durchgeführt. Wird es von Auftraggeberseite gewünscht, so kann der Anteil der kontrollierten Nettointerviews auch auf deutlich über 30% Prozent erhöht werden.

Der Fragebogen wird von Ipsos aus den einzelnen Themenblöcken zusammengestellt, wobei die jeweiligen Themenblöcke selbstverständlich in sich geschlossen bleiben. Jeder „Mitfahrer" erhält seine Fragen als Auszug aus dem Fragenprogramm. Außerdem erhalten alle an der jeweiligen Welle des Ipsos SOWI-Busses beteiligten Auftraggeber nach Feldbeginn eine Übersicht und Abfolge über die erhobenen Themen des Fragenprogramms – wobei aus Gründen des Urheberschutzes weder der Wortlaut der einzelnen Fragen noch die Anordnung dieser Fragen innerhalb eines Themenblocks den anderen Beteiligten mitgeteilt werden. Auf diese Weise erhalten alle Beteiligten einen Einblick, innerhalb welchen Umfelds und an welcher Position im Fragenprogramm die jeweiligen Frageblöcke stehen. Mit der Annahme des Angebots müssen sich allerdings die Auftraggeber damit einverstanden erklären, dass die Themenbereiche der jeweiligen Buswelle – ohne Identifizierung der Auftraggeber – allen Busteilnehmern zum Zwecke der Transparenz mitgeteilt wird.

3.2.1 Pretests

Vor Beginn jeder Erhebungswelle wird ein Pretest durchgeführt. Ziel eines solchen Pretests ist die Überprüfung und gegebenenfalls Modifizierung der Erhebungsinstrumente. Dazu werden insgesamt 30 Interviews (20 West, 10 Ost) unter Feldbedingungen, allerdings als Quotenstichprobe, durchgeführt. Die Interviewer senden die Frageprogramme zusammen mit einem Pretestbericht zurück, der sowohl qualitative Aussagen zu einzelnen Punkten enthält sowie einen standardisierten Teil mit geschlossenen Fragen zur Interviewdurchführung. Die Ergebnisse dieses Teils können mit bereits vorliegenden Ergebnissen aus unserer Pretest-Datenbank verglichen werden und erlauben somit eine objektive Beurteilung des Erhebungsinstruments hinsichtlich der erhobenen Kriterien.

Optional werden eine Reihe von weiteren Verfahren angeboten, die sich dann nur auf den jeweils beauftragten Frageblock beziehen. Dies können z.B. kombinierte Pretests aus teilnehmender Beobachtung und Interviews unter Feldbedingungen sein, „behavior coding" durch die Interviewer auf der Grundlage des

„klassischen“ Pretests, oder „Fokus-Gruppen“, bei denen sowohl einzelne Aspekte des Themas weiter exploriert werden, wie auch bestehende Instrumente durch die Befragten selber bewertet und ggf. verbessert werden. Da die Befragten der Pretests auf der Grundlage von Quoten rekrutiert werden, lassen sich hier bei Bedarf auch spezielle Zielgruppen definieren, über die der Fragebogen getestet werden soll.

Für jeden Pretest wird ein Methodenbericht angefertigt, in dem Vorkommnisse, Probleme und Lösungsvorschläge aufgeführt sind; dieser Methodenbericht wird mit dem Auftraggeber diskutiert.

Da durch den Einsatz von Computern die Antworten sofort erfasst werden, lässt sich eine deutlich höhere Datenqualität bei wesentlich größerem Gestaltungsspielraum des Fragenprogramms realisieren. So kann durch eine geeignete Programmierung bei widersprüchlichen Angaben durch den Interviewer sofort nachgefragt und geklärt werden, Fehleingaben können durch die Begrenzung der Wertefelder verhindert werden. Darüber hinaus lassen sich auch komplexe Filterführungen, Rotationen von Frageitems und –blöcken sowie Zufallsauswahlen von Items oder Fragen realisieren, die bei klassischen Paper & Pencil-Studien nicht möglich sind.

3.3 Feldarbeit der ersten Erhebungswelle

Die erste Erhebungswelle des Ipsos SOWI-Bus fand im Sommer 2005 in der Zeit vom 04. Juni bis 18. Juli statt. Dabei wurden 1.033 Interviews in der Hauptbefragung realisiert, zusätzlich sollte eine Aufstockung von 200 Interviews erfolgen, aus der 222 Interviews erzielt wurden. Die Gesamtfallzahl betrug insgesamt 1.255 Interviews. Für die Hauptbefragung waren 258 Sample-Points, für die Aufstockung 49 Sample-Points gezogen worden. Davon konnten jeweils 240 bzw. 41 auch mit Interviewerfolg bearbeitet werden. Die verbleibenden 18 Sample-Points konnten auch nach mehrfachem Einsatz der Interviewer nicht gefüllt werden. Möglicherweise spielte hier der später als ursprünglich geplant gelegene Feldbeginn eine Rolle, denn während der Feldzeit begannen bei acht Bundesländern bereits die Sommerferien. Insgesamt mussten 19 Sample-Points ersetzt werden, damit wurden 7,4% anstatt der maximal vorgesehenen 5% ersetzt. Auch hierbei spielt vermutlich der Beginn der Urlaubssaison eine Rolle, so dass einsetzbare Interviewer auch in geringerem Maße verfügbar waren. Umgekehrt mussten von der Feldleitung über die jeweils acht Adressen keine weiteren Adressen zugegeben werden. 32% der Interviews wurden mit Ergebnis kontrolliert, es mussten keine Interviews wegen Verdachts auf Unregelmäßigkeiten in der Durchführung entfernt werden.

Da der Stichprobenplan kein sog. „offenes Random Route“ war, also die Interviewer nicht so lange die Random-Route gingen, bis die erforderlichen acht Inter-

views erfüllt waren, sondern die Zahl der Adressen auf maximal acht beschränkt war, kann durch die Feldleitung auch die Bruttozahl der herausgegebenen Adressen genau bestimmt werden. Bei dieser Studie wurden für die Hauptbefragung insgesamt 1.806 Adressen ins Feld gegeben. Dies sind weniger als die rechnerisch möglichen 240*8=1.920 Adressen, ist aber durch die Feldsteuerung bedingt: Um den Feldverlauf besser kontrollieren zu können und eine Übererfüllung der angestrebten Nettostichprobe von 1.000 Interviews zu vermeiden, wurde die gesamte Feldzeit in insgesamt sechs Felder unterteilt, bei denen jeweils nur ein Teil der insgesamt erforderlichen Adressen eingesetzt wurde. Insofern konnte bei Erreichen der erforderlichen Fallzahl der Einsatz beendet werden, ohne dass alle 1.920 möglichen Adressen hätten eingesetzt werden müssen. Dies erklärt u.a. auch, warum keine Zusatzadressen, die möglich gewesen wären, eingesetzt werden mussten.

Daraus ergibt sich die folgende Ausschöpfung:

Tabelle 1: Ausschöpfung der ersten Ipsos SOWI-Bus Welle

1		*Bruttostichprobe (benutzbare Adressen)*	*1806*	
2		Stichprobenneutrale Ausfälle (ungültige Adresse wie Straße/Hausnummer nicht auffindbar, Wohnung unbewohnt, Anstaltshaushalt, sonstiges)	146	8,1%
3		*Nettostichprobe (1-2)* *davon:*	*1660*	*100,0%*
	3.1	Im HH mehrfach niemand angetroffen	188	11,3%
	3.2	HH verweigert jede Auskunft	184	11,1%
	3.3	ZP mehrfach nicht angetroffen	60	3,6%
	3.4	ZP verweigert Interview	162	9,8%
	3.5	Sonstige Befragtengründe (ZP spricht nicht deutsch, krank, sonstiges)	12	0,7%
4		Summe systematische Ausfälle	606	36,5%
5		Realisierte Interviews	1054	63,5%
6		Ausschöpfung (5/(4+5))	63,5%	

Die Ausschöpfung von 63,5% ist im Vergleich zu anderen Random-Stichproben durchaus gut. Sie spiegelt aber das bekannte Phänomen wider, dass Einwohnermeldeamtsstichproben, bei denen die Kontrolle der Adressen und zu befragenden Personen höher ist als bei einer Random-Stichprobe, eine (scheinbar) schlechtere Ausschöpfung erzielen (Koch 2002). Auch hier kann nicht völlig ausgeschlossen werden, dass in der Auflistung der Adressen zunächst eher „erfolgversprechen-

de“ Adressen in das Adressprotokoll eingetragen wurden, die dann mit dem entsprechenden Erfolg bearbeitet wurden, und somit die tatsächliche Ausschöpfung eher überschätzen. Dies bestätigt wiederum den Hinweis, dass eine Betrachtung alleine der Ausschöpfungsquote zur Beurteilung der Qualität von Umfragen durchaus problematisch ist.

3.4 Ipsos SOWI-Bus im Vergleich: ALLBUS 2004 und Demographische Standards 2004

Wenn sich die Qualität einer Stichprobe beweisen muss, so ist dies im Vergleich der Ergebnisse mit den bekannten Verteilungen der Grundgesamtheit sowie anderen Referenzstudien. Dazu werden im Folgenden einige zentrale demografische Merkmale (Alter, Geschlecht, Schulabschluss, Familienstand etc.) herangezogen und mit den Daten des ALLBUS 2004 sowie den vom Statistischen Bundesamt herausgegebenen „Demographischen Standards“ verglichen. Diese stellen eine Sonderauszählung des Mikrozensus 2003 dar, also einer – wenn auch amtlichen – Stichprobe aus der Grundgesamtheit der Bevölkerung. Ein solcher Vergleich über die Randverteilungen unterliegt natürlich einem ähnlichen Vorbehalt, wie er gegenüber Quotenstichproben formuliert wurde: Die Tatsache, dass die Verteilungen in einer Variable mehr oder weniger gut übereinstimmen, lässt keinen Schluss darüber zu, dass dies auch für andere, nicht beobachtete Verteilungen oder für die Strukturen gilt. Trotzdem ist dieser Vergleich der jeweiligen Verteilungen beider Erhebungen an einer gemeinsamen „Messlatte“ eine erste, auch intuitiv einleuchtende, Herangehensweise, und keinesfalls ungewöhnlich, wie z.B. die Analysen von Blohm et al. (2003) zum ALLBUS 2002 zeigen. Interne Qualitätskontrollen, wie bspw. der von Wolfgang Sodeur (1997) vorgeschlagene Abgleich zweier Variablen auf interne Konsistenz, müssen weiteren Vergleichen vorbehalten bleiben.

Beide Datensätze, SOWI-Bus wie ALLBUS, zielen als Grundgesamtheit auf die deutschsprachige Wohnbevölkerung der Bundesrepublik ab 18 Jahren ab, und sind von daher auch direkt miteinander vergleichbar.[2]

3.4.1 Alter und Geschlecht

Die korrekte Abbildung der Altersverteilung ist einer der wesentlichsten Prüfsteine für die Qualität einer Stichprobe. Alter und Geschlecht sind diejenigen erworbenen Merkmale, die mit der größten Messschärfe zu erheben sind, während andere zugeschriebene Merkmale (wie Status, Beruf) häufig Unschärfen aufweisen.

2 Da der ALLBUS eine disproportionale Stichprobe mit einer Überrepräsentation der neuen Bundesländer zieht, wurde diese Disproportionalität mit der Gewichtung über die Variable V891 wieder rückgängig gemacht.

Tabelle 2a: Altersverteilungen nach Geschlecht und Erhebungsinstrument

Alter	SOWI-Bus		ALLBUS 2004		Stat. BA 2004	
	Männer	Frauen	Männer	Frauen	Männer	Frauen
18 – 29	16,1	16,2	18,4	15,3	17,7	16,1
30 – 44	25,6	31,1	28,8	30,7	32,1	28,1
45 – 59	26,2	24,5	25,3	23,1	24,2	22,5
60 – 74	24,1	21,0	22,5	21,9	20,1	21,0
75 und älter	8,0	7,2	5,0	8,9	5,9	12,3

Quelle Stat. BA: Statistisches Jahrbuch 2004

Um die Daten differenzierter beurteilen zu können, wurde die Altersverteilung nochmals in Bezug auf Geschlecht differenziert (Tabelle 2a). Vergleicht man die Abweichungen je Zelle zwischen den jeweiligen Umfragen sowie zwischen den Stichproben und der Grundgesamtheit, wie sie vom Statistischen Bundesamt dargestellt wird (Tabelle 2b), so geben beide Stichproben die Verteilung der Grundgesamtheit im Prinzip recht gut wider. So entsprechen im SOWI-Bus die Verteilungen der 18-29jährigen und 60-74jährigen Frauen genau denen der Grundgesamtheit, mit etwas größeren Abweichungen (unter 1 Prozentpunkt) trifft dies auch der ALLBUS bei den 18-29jährigen, sowie 45-59- und 60-74jährigen Frauen und 75jährigen und älteren Männern. Größere Abweichungen (3 Prozentpunkte und mehr) finden sich in beiden Stichproben bei den 30-44jährigen Männern sowie den 75jährigen und älteren Frauen. Unter den Männern ist hier ein Selektionseffekt durch außerhäusliche Mobilität auf Grund von Berufstätigkeit zu vermuten, während bei den älteren Frauen dies eher auf ein Unsicherheitsgefühl gegenüber fremden Interviewern zurück zu führen sein dürfte. Diese Abweichungen fallen bei einem Random-Route-Verfahren erwartungsgemäß höher aus, da hier kaum die Möglichkeit besteht, z.B. durch ein Ankündigungsschreiben bestehende Befürchtungen zu zerstreuen und somit eine höhere Teilnahmebereitschaft zu erzielen. Darüber hinaus ist auch in der Random-Stichprobe unter den 60-74jährigen Männern eine höhere Abweichung gegenüber der Grundgesamtheit fest zu stellen, der auch ein erhöhter Wert in der Einwohnermeldeamtstichprobe entspricht.

Tabelle 2b: Differenzen der Altersverteilungen nach Geschlecht und Erhebungsinstrument

Alter	SOWI minus Stat. BA		ALLBUS minus Stat. BA	
	Männer	Frauen	Männer	Frauen
18 – 29	-1,6	0,1	0,7	-0,8
30 – 44	**-6,5**	3,0	**-3,3**	2,6
45 – 59	2,0	2,0	1,1	0,6
60 – 74	**4,0**	0,0	2,4	0,9
75 und älter	2,1	**-5,1**	-0,9	**-3,4**

3.4.2 Schulabschluss

Der höchste erreichte Schulabschluss ist als soziodemografischer Indikator sicherlich einer der ergiebigsten und in der Analyse am vielfältigsten einsetzbaren. In einem weitgehend über staatliche Zertifikate organisierten und strukturierten Schul- und Bildungssystem eröffnen bzw. schließen erreichte Schulabschlüsse Bildungswege auf bzw. ab, die später die Auswahl von Berufs- und Karrierechancen bestimmen. Der Schulabschluss ist immer auch eine erklärende Variable, wenn es um Fragen von Status, Prestige und sozialer Ungleichheit geht. Insofern ist eine verlässliche Messung dieser Anteile über die Qualität der Stichprobenrealisierung hinaus auch eine Voraussetzung für alle folgenden Analysen.

Betrachtet man die Realisierung der Verteilungen in den verschiedenen Zellen, zeigt sich, dass der Ipsos SOWI-Bus hier eine sehr gute Abbildung der Verteilung innerhalb der Grundgesamtheit liefert und kaum den Verteilungen des ALLBUS nachsteht. Beide Stichproben treffen die Verteilung derjenigen, die zum Zeitpunkt der Befragung noch Schüler sind, fast punktgenau, das gleiche gilt für die Verteilung derjenigen, die einen anderen als den vorgegebenen haben. Darüber hinaus weicht der SOWI-Bus bei den Anteilen der Fachhochschulabsolventen sowie derjenigen mit Hochschulreife etwas weniger von der Grundgesamtheit ab als der ALLBUS. Beide Stichprobenverfahren unterschätzen jedoch den Anteil derjenigen mit Volks- bzw. Hauptschulabschluss, und beide überschätzen die Gruppe derjenigen, die die Mittlere Reife haben, wobei das Random-Route-Verfahren hier die Verteilung in der Grundgesamtheit noch etwas stärker verfehlt als die Einwohnermeldeamtstichprobe.

Tabelle 3: Verteilungen und Differenzen der Verteilungen der Schulabschlüsse nach Erhebungsinstrument

Schulabschluss	SOWI-Bus	ALLBUS 2004	Demogr. Standards 2004	SOWI-Bus minus Demogr. Stand.	ALLBUS minus Demogr. Stand.
Noch Schüler	1,4	1,0	1,1	0,3	-0,1
Ohne Abschluss	3,4	2,6	-	-	-
Volks- / Hauptschule	41,0	41,8	43,5	-2,5	-1,7
Mittlere Reife	32,8	30,8	26,9	**5,9**	**3,9**
Fachhochschulreife	5,2	6,3	4,8	0,4	1,5
Hochschulreife	15,6	17,2	16,2	-0,6	1,0
Anderer Abschluss	0,6	0,3	0,9	-0,3	-0,6

3.4.3 Soziale Stellung

Die soziale Stellung ist eine zentrale Variable in vielen Analysen, weil sie mit einer Vielzahl von Variablen zusammenhängt und diese fast wie in einem Brennglas bündelt. Sie informiert u.a. darüber, mit welchen Chancen und Möglichkeiten die Inhaber einer solchen Stellung rechnen können bzw. wie weit sie auch von bestimmten Chancen und Möglichkeiten ausgeschlossen sind. In die Zuordnung zu einer bestimmten Stellung gehen das Lebensalter ein – z.B. im Vergleich von Schülern und Studenten vs. Rentnern und Pensionären – und damit Lebenserfahrungen und berufliche und persönliche Karrieremöglichkeiten. Arbeitslose oder Menschen, die auf Null- oder Kurzarbeit gesetzt wurden, haben auf Grund der finanziellen Einschränkungen nur noch beschränkte Möglichkeiten zur gesellschaftlichen Teilhabe, darüber hinaus erleben sie in der Regel zwangsweise Brüche in ihrer persönlichen Berufsbiografie, die häufig auch psychische Brüche umfassen. Hausfrauen und –männer leben in einer anderen Alltagswelt als Erwerbstätige, mit anderen Anforderungen, Kontaktmöglichkeiten und persönlichen Entwicklungsmöglichkeiten. Eine korrekte Erfassung der sozialen Stellung ist daher – wie auch Schulabschluss oder Erwerbstätigkeit – nicht nur für eine sozialstrukturelle Beschreibung, sondern eben auch als unabhängige Variable für weitere Analysen unerlässlich.

Tabelle 4: Verteilungen und Differenzen der Verteilungen der sozialen Stellung nach Erhebungsinstrument

Soziale Stellung	SOWI-Bus	ALLBUS 2004	Demogr. Standards 2004	SOWI-Bus minus Demogr. Stand.	ALLBUS minus Demogr. Stand.
Schüler/Student	5,5	5,0	2,6	2,9	2,4
Rentner/Pensionär	27,7	25,0	29,0	-1,3	**-4,0**
Zur Zeit arbeitslos, Null- / Kurzarbeit	10,8	6,3	5,1	**5,7**	1,2
Hausfrau/Hausmann	10,6	11,0	8,5	2,1	2,5
Wehr- / Zivildienstleistender	0,0	0,2	0,2	-0,2	0,0
Aus anderen Gründen nicht erwerbstätig	2,8	3,5	11,7	**-8,9**	**-8,2**
Übrige	42,5	48,8	42,8	-0,3	**6,0**

Beide Stichprobensysteme bilden die verschiedenen Gruppen weitgehend ähnlich gut ab, auch dort, wo es Abweichungen gibt, gehen diese weitgehend in dieselbe Richtung. Fast punktgenau wird der Anteil der Wehr- und Zivildienstleistenden in beiden Stichproben abgebildet, beide überschätzen etwas sowohl den Anteil der Schüler und Studenten sowie den Anteil der Hausfrauen bzw. –männer. Deutlich unterschätzt, und zwar in fast derselben Größenordnung, wird der Anteil derjenigen, die aus anderen Gründen nicht erwerbstätig sind. Die Abweichung von -8,9 (SOWI-Bus) bzw. -8,2 (ALLBUS) Prozentpunkten ist dabei die größte hier beobachtete Abweichung. Möglicherweise liegen hier definitorische Unschärfen zwischen den Erhebungsinstrumenten der beiden Surveys einerseits und der Berechnung aus den demografischen Standards andererseits vor. Während der ALLBUS den Anteil der Residualkategorie überschätzt und den Anteil der Rentner und Pensionäre unterschätzt, wird im Ipsos SOWI-Bus der Anteil der Arbeitslosen, Null- und Kurzarbeiter überschätzt. Dabei ist aber zu berücksichtigen, dass sich der SOWI-Bus auf den Zeitpunkt der Erhebung im Frühsommer 2005, während sich der ALLBUS auf das Jahr 2004 bezieht. Gerade der Arbeitslosenanteil ist jedoch in den letzten Jahren deutlich gestiegen und darüber hinaus noch einer starken saisonalen Schwankung unterworfen.

3.4.4 Erwerbstätigkeit

Fast komplementär zur Frage nach der sozialen Stellung ist die Erfassung der Erwerbstätigkeit. Während die Erwerbstätigen als Residualkategorie in der Erfassung der sozialen Stellung enthalten sind, werden sie in dieser Frage noch nach

den jeweiligen Definitionen der Vollzeit-, Teilzeit- und Nebenerwerbstätigkeit unterschieden. Hier zeigen sich in den Stichproben allerdings deutliche Unterschiede in der Erfassung der jeweiligen Kategorien. So weist der Ipsos SOWI-Bus nur eine Abweichung für die Teilzeiterwerbstätigen von 0,3 Prozentpunkten auf, während der ALLBUS diese mit -2,3 Prozentpunkten unterschätzt. Gleichzeitig unterschätzen beide den Anteil der Vollzeiterwerbstätigen, und beide überschätzen den Anteil der Nebenerwerbstätigen. Dies kann als Hinweis darauf gewertet werden, dass sowohl Vollzeitberufstätige weniger bereit sind, an Interviews teilzunehmen, wie auch dass Interviewer bei diesen schwerer Erreichbaren auf eher leichter erreichbare Personen – wie eben Nebenerwerbstätige – ausweichen. So unterschätzt der SOWI-Bus die Vollzeiterwerbstätigen um 5,9 Prozentpunkte, während dies beim ALLBUS lediglich 1,3 Prozentpunkte sind. Umgekehrt überschätzt der SOWI-Bus die Nebenerwerbstätigen um 5,5 Prozentpunkte, während der ALLBUS hier eine um immer noch 3,5 Prozentpunkte zu hohe Verteilung ausweist. Dies unerwünschte, wenn auch bekannte, Verhalten ist designbedingt stärker bei Random-Route-Verfahren als bei Einwohnermeldeamtsstichproben, da – wie oben gezeigt – Random-Route-Verfahren mehr Freiheitsgerade für Interviewer beinhalten als Personenstichproben aus den Einwohnermeldeamtsregistern.

Tabelle 5: Verteilungen und Differenzen der Erwerbstätigkeit nach Erhebungsinstrument

Erwerbstätigkeit	SOWI-Bus	ALLBUS 2004	Demogr. Standards 2004	SOWI-Bus minus Demogr. Stand.	ALLBUS minus Demogr. Stand.
Vollzeit	70,2	74,8	76,1	**-5,9**	-1,3
Teilzeit	16,9	14,3	16,6	0,3	-2,3
Nebenher erwerbstätig	12,9	10,9	7,4	**5,5**	**3,5**

3.4.5 Familienstand

Waren bei den bisher betrachteten Variablen die Unterschiede in den Verteilungen noch relativ gut erklärbar, zeigt sich in der Verteilung des Familienstandes eine Zelle, die deutlich von der Verteilung in der Grundgesamtheit abweicht, für die es aber keine schlüssige Erklärung gibt. Trotzdem soll – der Redlichkeit halber – auch diese hier dargestellt werden. So hat der Ipsos SOWI-Bus den Anteil der Geschiedenen um 7 Prozentpunkte gegenüber der amtlichen Statistik zu hoch geschätzt, damit wird der Anteil der Verheirateten wie der Ledigen entsprechend zu niedrig geschätzt. Da das Merkmal „geschieden“ keine erkennbare Auswir-

kung auf die Zielpersonen- oder Haushaltsauswahl hat, kann ein durch das Stichprobendesign bedingter Fehler ausgeschlossen werden. Ein, diesen Effekt aber auch nicht vollständig erklärender, Grund könnte in der den Befragten vorgelesenen Definition von „Ledig“ liegen: Durch den Nachsatz „Ledig, nie verheiratet“ (der beim ALLBUS fehlt) können diejenigen Geschiedenen, die sich ansonsten als „ledig“ betrachtet hätten, dann tatsächlich „geschieden“ angegeben haben. Dies kann aber höchstens die Unterschätzung der Ledigen mit einem gewissen Kompensationseffekt zu Gunsten der Geschiedenen erklären, aber auch nicht die Unterschätzung der Verheirateten. Da Codierfehler ausgeschlossen werden können, kann hier als Erklärung dafür letztlich nur ein zufälliger Ausreißer, wie er durch Zufallsstichproben immer auch gegeben sein kann, dienen. In der folgenden SOWI-Bus Erhebung wird die Hypothese des Ausreißers dann zu überprüfen sein.

Tabelle 6: Verteilungen und Differenzen der Verteilungen des Familienstandes nach Erhebungsinstrument

Familienstand	SOWI-Bus	ALLBUS 2004	Demogr. Standards 2004	SOWI-Bus minus Demogr. Stand.	ALLBUS minus Demogr. Stand.
Verheiratet, mit Partner	51,6	60,7			
Verheiratet, getrennt	2,0	1,5			
Verheiratet, gesamt	53,6	62,2	58,9	**-5,3**	**3,3**
Verwitwet	10,9	7,3	9,5	1,4	-2,2
Geschieden	13,8	6,7	6,8	**7,0**	-0,1
Ledig	21,7	23,8	24,9	**-3,2**	-1,1

3.4.6 Zusammenfassung

Insgesamt zeigt dieser Vergleich zwischen den Ergebnissen der ersten Welle des Ipsos SOWI-Bus, dem ALLBUS als sozialwissenschaftlicher Referenzstudie, und den Grundgesamtheiten, wie sie von der offiziellen Statistik ausgewiesen werden, dass auch eine sorgfältig durchgeführte Random-Walk-Studie zuverlässige und genaue Ergebnisse produziert. In der Erfassung zentraler Variablen wie Alter, Geschlecht, Schulbildung oder sozialer Stellung liefern beide Stichprobensysteme in der Genauigkeit vergleichbare Ergebnisse, in der Erfassung der Erwerbstätigkeit ist der ALLBUS (wie zu erwarten) genauer. Insofern (und zunächst auch nur für die hier berichteten Variablen) ist tatsächlich zu fragen, ob der zusätzliche Aufwand, der bei einer Einwohnermelderegisterstichprobe betrieben

werden muss, die zusätzliche Präzision der Ergebnisse rechtfertigt. Als Referenzstudie ist jedoch die Vorgabe eines theoretisch möglichst präzisen Stichprobensystems unabdingbar, wie groß die damit gewonnene zusätzliche Präzision ist, ist demnach eine eher empirische Frage.

Für Studien wie den Ipsos SOWI-Bus ist dieses Ergebnis umgekehrt eine Bestätigung, dass es nicht unbedingt der extrem aufwändigen Studiendesigns bedarf, um Daten hoher Präzision zu erbringen, sondern dass bereits mit sorgfältig durchgeführten Random-Walk-Designs zuverlässige Ergebnisse erzielbar sind. Insofern kann dies auch als eine Bestätigung der These gesehen werden, dass es tatsächlich mehrere korrekte Verfahren gibt, aus denen innerhalb eines zeitlichen und budgetären Rahmens die sozialwissenschaftlichen Auftraggeber in Abhängigkeit von ihrem Erkenntnisinteresse die jeweils optimale Kombination wählen können.

4 Zusammenfassung

Für die Akzeptanz von Befragungsergebnissen in der wissenschaftlichen wie allgemeinen Öffentlichkeit ist der Nachweis von Qualität entscheidend. Über die Definition dessen, was Qualität in Bezug auf sozialwissenschaftliche Erhebungen jedoch ist, wird noch trefflich gestritten. Allein über die Kriterien „richtige Ergebnisse“ und „korrekte Verfahren“ wird ein Qualitätsnachweis nicht erfolgreich sein, da die Profession mittlerweile eine ganze Reihe korrekter Verfahren zur Verfügung stellt. An dieser Stelle wird statt dessen argumentiert, dass die Qualität einer Studie danach zu bemessen ist, ob die verwendeten Verfahren angemessen und geeignet sind, ein bestimmtes Erkenntnisinteresse zu verfolgen. Über die unterschiedlichen Erkenntnisinteressen ergeben sich dann jeweils unterschiedliche Anforderungen, die sich anhand der Dimensionen angestrebte Genauigkeit, verfügbare Zeit und verfügbares Budget einordnen lassen. Die gängigsten zur Verfügung stehenden Stichprobendesigns wurden anschließend kurz beschrieben und in Bezug auf ihre Qualitätsmerkmale eingeordnet. In einen zweiten Teil wurde der Ipsos SOWI-Bus als sozialwissenschaftliche Mehrthemenumfrage vorgestellt, der qualitativ im oberen Feld der Random-Walk-Verfahren angesiedelt ist. In einem Vergleich einiger zentraler Variablen mit dem ALLBUS wurde die Qualität der Stichprobe des Ipsos SOWI-Bus überprüft. Dabei zeigt sich, dass für eine Reihe von Variablen die Genauigkeit und damit Qualität der beiden Stichprobensysteme durchaus vergleichbar ist.

5 Literatur

ADM (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.) (1979): Musterstichprobenpläne. München: Verlag Moderne Industrie.

ADM (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.) / ASI (Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V.) / BVM (Berufsverband Deutscher Markt- und Sozialforscher e.V.) (1999): Standards zur Qualitätssicherung in der Markt- und Sozialforschung. Frankfurt/M.: ADM.

ADM (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.) / AG.MA (Arbeitsgemeinschaft Media-Analyse e.V.) (1999): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis. Opladen: Leske & Budrich.

Althoff, S. (1993): Auswahlverfahren in der Markt-, Meinungs- und empirischen Sozialforschung. Pfaffenweiler: Centaurus.

Biemer, P. P. / Lyberg, L. E. (2003): Introduction to Survey Quality. Hoboken: Wiley.

Blohm, M. / Harkness, J. / Klein, S. / Scholz, E. (2003): Konzeption und Anlage der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2002. ZUMA-Methodenbericht 2003/12. Überarbeitete Version, August 2004. Mannheim: ZUMA.

Cochran, W. G. (1972): Stichprobenverfahren. Berlin u.a.: de Gruyter.

Daves, R.P. / Newport, F. (2005) : Pollsters under Attack. 2004 Election Incivility and its Consequences. In: Public Opinion Quarterly, Vol. 69, No. 5, S. 670-681.

DFG (Deutsche Forschungsgemeinschaft) (1999): Qualitätskriterien der Umfrageforschung. Denkschrift. Berlin: Akademie Verlag.

Faas, T. (2003): Umfragen im Umfeld der Bundestagswahl 2002: Offline und Online im Vergleich. In: ZA-Information, H. 52, S. 120-134.

Gabriel, O.W. / Keil, S. I. (2005): Empirische Wahlforschung in Deutschland: Kritik und Entwicklungsperspektiven. In: Falter, J. W. / Schoen, H. (Hrsg.): Handbuch Wahlforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 661-641.

Hoffmeyer-Zlotnik, J.H.P. (1997): Sozialwissenschaften-Bus 1998. In: ZUMA-Nachrichten Jg. 21, H. 41, S. 189-191.

Gabler, S. / Häder, S. / Hoffmeyer-Zlotnik, J.H.P. (Hrsg.) (1998): Telefonstichproben in Deutschland. Wiesbaden: Westdeutscher Verlag.

Koch, A. (2002): 20 Jahre Feldarbeit im ALLBUS: Ein Blick in die Blackbox. In: ZUMA-Nachrichten Jg. 26, H. 51, S. 9-37.

Mohler, P. Ph. / Koch, A. / Gabler, S. (2003): Alles Zufall oder? Ein Diskussionsbeitrag zur Qualität von face-to-face-Umfragen in Deutschland. In: ZUMA-Nachrichten Jg. 27, H. 53, S. 10-15.

Neller, K. (2005): Kooperation und Verweigerung: Eine Non-Response-Studie. In: ZUMA-Nachrichten Jg. 29, H. 57, S. 9-36.

Quatember, A. (2001): Die Quotenverfahren. Stichprobentheorie und -praxis. Aachen: Shaker Verlag.

Schneekloth, U. / Leven, I. (2003): Woran bemisst sich eine „gute" Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. In: ZUMA-Nachrichten Jg. 27, H. 53, S. 16-57.

Schnell, R. / Hill, P. B. / Esser, E. (1999): Methoden der empirischen Sozialforschung. München u.a.: Oldenbourg.

Sodeur, W. (1997): Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. In: ZA-Information Jg. 19, H. 41, S. 58-82.

Statistisches Bundesamt (2004): Demographische Standards. Eine gemeinsame Empfehlung des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute e.V. (ADM), der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI) und des Statistischen Bundesamtes. Wiesbaden: Statistisches Bundesamt.

v. Harder, B. / Hoffmeyer-Zlotnik, J.H.P.: Der Sozialwissenschaft-Bus zukünftig auch in der DDR. In: ZUMA-Nachrichten Jg. 14, H. 26, S. 79-81.

Design und Schätzqualität im registergestützten Zensus
Ergebnisse einer Monte-Carlo-Studie

Ralf Münnich, Kersten Magg

Zusammenfassung

2011 soll in Deutschland der nächste Zensus stattfinden. Im Gegensatz zur klassischen Vollerhebung soll ein registergestützter Zensus durchgeführt werden. Ziel ist es, mit Hilfe einer zusätzlichen Stichprobe die möglichen Fehlbestände und Karteileichen in den Registern zu schätzen und somit korrigierte Zensuswerte auszuweisen. Mögliche weitere interessierende Variablen, die nicht aus einem Registerabzug ermittelt werden können, sollen auf Basis geeigneter Hochrechnungsverfahren geschätzt werden. Zur Diskussion stehen sowohl klassische Schätzverfahren als auch die moderneren Verfahren der Small Area-Schätzung.

Ziel der Untersuchung ist es, mit Hilfe einer Monte-Carlo-Studie auf Basis einer speziell errichteten Gesamtheit erste Ergebnisse für die Auswirkung verschiedener Stichprobendesigns auf die Ermittlung des Bevölkerungsumfangs inklusive interessierender Teilgruppen sowie der Schätzung zusätzlicher interessierender Variablen zu erhalten.

Stichworte: Zensus, Register, Design, Schätzqualität, Small Area-Schätzung

1 Einführung in die Problematik

Der für das Jahr 2011 in Deutschland erstmals vorgesehene registergestützte Zensus verlangt einen statistischen Paradigmenwechsel und damit ein teilweises Umdenken der Nutzer der Zensusdaten. Im Gegensatz zum klassischen Zensus (Vollerhebung) – der indes auch nicht als fehlerfrei angesehen werden darf – basiert der registergestützte Zensus zunächst auf Daten der Einwohnermeldeämter, deren Datenbasis durch eine Ergänzungsstichprobe erweitert wird. Dabei sind zwei wesentliche Ziele zu berücksichtigen:

- Auszählung des Bevölkerungsumfangs, untergliedert nach geographischen und demographischen Merkmalen;
- Gewinnung weiterer ökonomischer und soziodemographischer Daten.

Der erste Punkt steht sicherlich zunächst im Vordergrund einer amtlichen Statistik. Konkret dient in Deutschland eine derartige Erhebung zur Feststellung des Status quo und zur Adjustierung der Eckdaten der Bevölkerungsfortschreibung. Bei der Verwendung der Einwohnermelderegister für eine derartige Auszählung müssen vorrangig zwei Fehlerquellen berücksichtigt werden. Zum einen Personen, die zwar gemeldet, aber nicht mehr vorhanden sind, die so genannten Karteileichen, und zum anderen Personen, die vorhanden, aber nicht gemeldet sind, die so genannten Fehlbestände. Beide Fehlerquellen müssen mit Hilfe zusätzlicher Stichprobeninformationen abgeschätzt werden, um korrigierte Gesamtzahlen für den Bevölkerungsbestand ableiten zu können. Neben dem Bevölkerungsbestand interessieren zusätzlich noch Haushaltsinformationen, die im Allgemeinen im Einwohnermelderegister nur unzureichend vorhanden sind. Mit Hilfe der Daten der Einwohnermelderegister, den Abschätzungen der Anzahl der Karteileichen und Fehlbestände, sowie den Daten zu Gebäude- und Wohnungszählung werden schließlich Haushaltsinformationen gewonnen. Informationen zu dieser Haushaltegenerierung, den Ergebnissen zum Zensustest der amtlichen Statistik und eine eingehendere Erörterung der Problematik können in Statistisches Bundesamt (2004), Eppmann (2004), Schäfer (2004) sowie Magg et al. (2006) nachgelesen werden.

Der zweite Punkt befasst sich mit der Gewinnung zusätzlicher Informationen über soziodemographische und ökonomische Variablen. Im Gegensatz zur eigentlichen Zensus-Problematik, die als Basis die Daten der Einwohnermelderegister verwenden kann, müssen für die meisten zusätzlichen Variablen Schätzwerte auf Basis der Stichprobe gewonnen werden.

Da bisher in Deutschland noch kein registergestützter Zensus durchgeführt worden ist, sollte mit Hilfe einer einfachen beispielorientierten Monte-Carlo-Studie auf Basis einer synthetischen, aber realitätsnahen Gesamtheit Aufschluss über die Einsetzbarkeit ausgewählter Schätzverfahren im registergestützten Zensus gewonnen werden. Gegenstand dieser ersten Untersuchungen ist eine Analyse der wesentlichen Einflussfaktoren einer geeigneten Schätzung des Bevölkerungsbestandes sowie ausgewählter möglicher zusätzlicher Variablen. Dabei stehen folgende Fragen im Vordergrund:

- Welche Methoden können im Rahmen der verschiedenen Zielsetzungen herangezogen werden?
- Inwieweit wirken sich verschiedene mögliche Stichprobendesigns auf die Ergebnisse aus?
- Welche räumlichen und sachlichen Untergliederungen lassen überhaupt noch qualitativ akzeptable Ergebnisse zu?

Grundlage der Untersuchungen waren Überlegungen aus dem Zensustest, wie etwa das Stichprobendesign sowie die Verteilung von Karteileichen und Fehlbe-

ständen betreffend. Dabei sollte anhand von Simulationen auf Basis einer Beispielgesamtheit die Wirksamkeit ausgewählter Schätzverfahren in Bezug auf Gesamtwerte sowie auf Teilgesamtwerte überprüft werden.

Im nächsten Abschnitt werden zunächst ausgewählte Schätzverfahren vorgestellt. Anschließend werden die Daten, die der Simulationsstudie zugrunde lagen, dargestellt. Schließlich folgt eine eingehende Erörterung der interessierenden Fragestellungen und der ersten Ergebnisse der Studie sowie eine Zusammenfassung nebst Ausblick.

2 Klassische und Small Area-Schätzmethoden

2.1 Schätzung von Totalwerten und der Horvitz-Thompson-Schätzer

Die nachfolgend zu erörternden Schätzverfahren beruhen teilweise auf dem klassischen Horvitz-Thompson-Schätzer bzw. verwenden eine spezielle Darstellungsform eines Horvitz-Thompson-Schätzers.

Man verwendet zur Schätzung eines Totalwertes τ_Y des interessierenden Merkmals *Y* einer endlichen Population

$$(2.1) \quad \hat{\tau}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i \cdot y_i,$$

wobei die Designgewichte $d_i = 1/\pi_i$ die inversen Inklusionswahrscheinlichkeiten erster Ordnung sind. Mit Hilfe dieser Gewichte erhält man für verschiedene Stichprobendesigns eine erwartungstreue Schätzung für den Totalwert des interessierenden Merkmals τ_Y.

Der Horvitz-Thompson-Schätzer geht auf Horvitz und Thompson (1952) zurück. Eine eingehende Erörterung des Horvitz-Thompson-Schätzers sowie der zugehörigen Varianzschätzmethoden kann bspw. Särndal et al. (1992) bzw. Deville (1999) entnommen werden.

2.2 Kalibrierung und *g*-Gewichte

Bei gegebenem Stichprobendesign werden zur Verbesserung von Schätzungen gerne weitere verfügbare Informationen herangezogen. Von besonderem Interesse sind Informationen, für die neben der reinen Stichprobeninformation auch Totalwertinformationen aus der interessierenden Population vorliegen. In Haushaltsstichproben sind dies zumeist die Bevölkerungszahlen mit etwaigen Teilinformationen über Altersgruppen sowie Geschlecht und Nationalität. Mögliche

Disproportionen in der Stichprobe können dann mit Hilfe geeigneter Gewichtungen korrigiert werden.

Ausgehend vom Horvitz-Thompson-Schätzer (2.1) verwendet man dann den so genannten gewichteten Horvitz-Thompson-Schätzer

$$\hat{\tau} = \sum_{k \in S} \underbrace{w_k \cdot d_k}_{g_k} \cdot y_k$$

mit den Gewichten $w_k = w_k(\mathbf{x}_k)$, die auf möglichen verfügbaren Zusatzinformationen basieren können, und den zuvor eingeführten Designgewichten d_k. Der Gewichtevektor **g** umfasst beide Komponenten – auf ihn wird zumeist mit *g*-Gewichten rekurriert.

Im Rahmen von Kalibrierungsverfahren werden nun solche Gewichtungen untersucht, für die gerade eine Kalibrierung der Hilfsinformationen erreicht wird, also

$$\sum_{k \in S} g_k \cdot \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k \tag{2.2}$$

gilt. Innerhalb der zahlreichen möglichen Gewichtungen, die diese Nebenbedingungen erfüllen, wird diejenige gewählt, die im Sinne einer geeignet zu wählenden Distanzfunktion den geringsten Abstand zur Ausgangsgewichtung der Designgewichte aufweist. Diese sichert unter gewissen Regularitätsbedingungen günstige asymptotische Eigenschaften der Kalibrierungsschätzer (vgl. Deville, 1999).

Gesucht ist nun eine Funktion *G*, so dass das Funktional

$$\min_{w} \sum_{k \in S} d_k G\left(\frac{g_k}{d_k}\right) = \min_{w} \sum_{k \in S} d_k G(w_k)$$

unter der Nebenbedingung (2.2) zu minimieren ist. Dabei sollten nur solche Funktionen *G* verwendet werden, für die $G(1)=0$, $G'(1)=0$ bzw. $G''(1)=1$ gilt. Diese Bedingungen sichern eine Minimalität der Zielfunktion als Abstand zu den Designgewichten.

Drei bekannte Fälle werden bevorzugt herangezogen:

Raking	$G(u) = u \cdot \log(u) - u + 1$
ML-Raking	$G(u) = u - 1 - \log(u)$
GREG	$G(u) = \frac{1}{2}(u-1)^2$

Im dritten Fall resultiert der bekannte verallgemeinerte Regressionsschätzer (GREG), der damit ebenso die Kalibrierungseigenschaft (2.2) erfüllt. Für eine eingehende Diskussion der Kalibrierungsschätzer sei insbesondere auf die grundlegende Arbeit von Deville und Särndal (1992) sowie auf Deville (1999) und D'Arrigo und Skinner (2003) und die darin zitierten Arbeiten verwiesen. Eine Varianzschätzung der Kalibrierungsschätzer erfolgt zumeist mit Hilfe von Residualvarianzschätzern, weist jedoch ähnliche Probleme bei der Bestimmung der Inklusionswahrscheinlichkeiten zweiter Ordnung für allgemeinere Stichprobendesigns auf wie der Horvitz-Thompson-Schätzer.

Eine ausführliche Studie zum Vergleich der Effizienz der vorgestellten Verfahren unter Berücksichtigung von Nonresponse sowie dessen Kompensation kann Davison et al. (2004) entnommen werden. Dabei muss angemerkt werden, dass ein bemerkenswerter Unterschied zwischen den Ergebnissen der vorgestellten Verfahren nur sehr selten in Erscheinung tritt. Lediglich in Sonderfällen kann man unterschiedliche Ergebnisse beobachten. Vielmehr spielen zwei andere Aspekte hier eine größere Rolle. Im Gegensatz zum GREG-Schätzer, der direkt in Matrixform ermittelt werden kann, müssen bei den anderen Verfahren spezielle Algorithmen zur Anwendung kommen, welche im Einzelfall zu einer deutlichen Verlangsamung der Berechnungen führen können. Dafür kann es beim GREG-Schätzer passieren, dass negative Gewichte resultieren. Dies mag im Sinne eines Gesamtwertes effizient und damit auch vertretbar sein. Da in der amtlichen Statistik solche Gewichtungsvektoren jedoch Nutzerdaten beigefügt werden müssen, können Sonderauswertungen schnell auch zu unplausiblen und damit nicht vertretbaren Ergebnissen führen. Hier erweist sich der Raking-Schätzer als eher geeignet, da bei ihm stets positive Gewichte resultieren.

Im Rahmen der Zensus-Problematik entsteht indes noch eine weitere Schwierigkeit, nämlich inwieweit Informationen aus übergeordneten Populationen zur Kalibrierung herangezogen werden. Möchte man etwa auf Gemeindeebene Schätzungen durchführen, die auf Grund möglicherweise sehr geringer Stichprobenumfänge hohe Standardfehler aufweisen können, dann eignet sich zumeist eine Verwendung der Hilfsinformationen auf Bezirks- oder auch Bundeslandsebene. Eine Darstellung des so genannten Domain-spezifischen Regressionsschätzers kann Särndal et al. (1992), S. 386 ff., entnommen werden.

2.3 Grundmodell der Small Area-Schätzung

Neben den klassischen Schätzmethoden werden spezielle Verfahren der Small Area-Methodik in die Untersuchung aufgenommen. Grundsätzlich handelt es sich dabei um Verfahren, welche sich durch ihre Eigenschaften besonders dafür eignen, Schätzwerte kleiner Subpopulationen zu generieren. Diese Subpopulationen können sowohl inhaltlich als auch geographisch abgegrenzt sein. Der Einsatz dieser Small Area-Verfahren empfiehlt sich generell dann, wenn direkte bzw. klassische Verfahren nicht in der Lage sind, zuverlässige Schätzwerte für diese Subpopulationen hervorzubringen. Einen umfassenden Überblick über die Methodik und Anwendungsmöglichkeiten der Small Area-Schätzung gibt Rao (2003).

Für die Testsimulationen wurde ein synthetischer Schätzer ausgewählt, dessen Ergebnisse mit denen der klassischen Verfahren verglichen werden sollen. Dieser Schätzer zählt zu den so genannten modellbasierten Schätzern. Diese charakterisieren sich dadurch, dass explizite Modellannahmen, welche einen erklärenden Beitrag zu den Unterschieden und der Variabilität zwischen den Small Areas liefern sollen, der Schätzwertberechnung zugrunde gelegt werden.

In den vorliegenden Untersuchungen basiert die Small Area-Schätzung auf dem linearen Regressionsmodell

$$(2.3) \quad y_{i,d} = x'_{i,d}\beta + u_d + e_{i,d} \, .$$

Dabei wird angenommen, dass die Hilfsinformation $x_{i,d}$ für das Individuum i in der Small Area d vorhanden ist und die interessierende Variable $y_{i,d}$ damit erklärt werden kann. Bei u_d und $e_{i,d}$ handelt es sich um den Area-spezifischen Effekt (bspw. auf Gemeindeniveau) mit $u_d \sim iid\ N(0;\sigma_u^2)$ sowie um den zufälligen Störterm mit $e_{i,d} \sim iid\ N(0;\sigma_e^2)$. Der synthetische Schätzer des so genannten *Standard two-level* – Modells (Schätzer des Modells A) ist schließlich für den Mittelwert des Zielparameters $\overline{Y}_d$ gegeben durch

$$(2.4) \quad \hat{\overline{Y}}_{d,\,SynthA} = \overline{X}'_d\hat{\beta},$$

wobei $\overline{X}_d$ die wahren Mittelwerte der Hilfsvariablen in Area d angibt.

Aufgrund der zugrundegelegten Modelle wird es möglich sein, Schätzwerte für Small Areas zu erhalten, für die nur wenig bzw. sogar keine Stichprobendaten vorliegen. Im Kontext des registergestützten Zensus handelt es sich dabei zum Beispiel um Schätzungen in kleinen Gemeinden oder Stadtteilen, welchen in der Stichprobe eine nur geringe Gewichtung zukommt. Aufgrund von Hilfsvariablen oder Zielvariablen aus anderen Small Areas ist es mit Hilfe des erklärten Zusam-

menhangs der Areas untereinander möglich, Schätzwerte zu generieren. In der Literatur unterscheidet man zahlreiche Arten von Modellen, die zur Erklärung herangezogen werden können. Wenngleich einerseits in der Vielzahl und Flexibilität der Modellierung ein großer Vorteil der Small Area-Schätzung zu sehen ist, stellt diese Tatsache gleichzeitig die Schwierigkeit dar, diese Methodik adäquat anzuwenden. Die Schätzung der Varianz sowie die damit verbundene Ermittlung von Maßen (Design-Effekt, Konfidenzintervallüberdeckungsraten) stellt sich in der Regel als schwierig dar. Anhand einer entsprechenden Modelldiagnostik können Modelle allerdings evaluiert und gegebenenfalls angepasst und verbessert werden, um optimierte Schätzwerte auf den Small Areas zu erhalten. Ein weiterer Vorteil dieser Modelle ist, dass Area-spezifische Präzisionsmaße ermittelt werden können, die ebenso zur Beurteilung des Schätzverfahrens herangezogen werden können. Häufig vorkommende Modellvariationen liegen in der beliebigen problemadäquaten Ergänzung um fixe und zufällige Effekte, die den Erklärungsgehalt detaillierter spezifizieren können sowie in der Anzahl der betrachteten Ebenen. In diesem Kontext spricht man von so genannten Mehrebenenmodellen oder multilevel-Modellen (vgl. Longford, 1993, und Goldstein, 1995). Im Zusammenhang mit modellbasierten Schätzverfahren spricht man in der Literatur von Empirical Best Linear Unbiased Predictor, Empirical Bayes Method oder Hierarchical Bayes Method. Die Charakteristika dieser Methoden liegen vor allem in der Art und Weise des Auffindens der Lösung der linearen bzw. nichtlinearen Gleichungssysteme sowie in speziellen Modellanpassungen und Skalierungsproblemen.

Weitere Informationen und Details zu modellbasierten Small Area-Methoden sind in Rao (2003) und Longford (2005) sowie in Münnich et al. (2004) zu finden. Außerdem bietet Magg et al. (2006) und die darin zitierte Literatur einen Überblick über unterschiedliche Ansätze der Small Area-Schätzung und die Eingliederung der modellbasierten Schätzer.

Small Area-Verfahren werden bereits seit vielen Jahren konzipiert, rege diskutiert und ständig weiterentwickelt sowie auf neue Anwendungen adaptiert. Für statistische Auswertungen hat die Small Area-Fragestellung schon seit Jahrzehnten stark an Bedeutung gewonnen. Bereits im 11. Jahrhundert soll es in England sowie im 17. Jahrhundert in Kanada erste derartige statistische Auswertungen gegeben haben (vgl. Ghosh und Rao, 1994, S. 55 und die darin zitierte Literatur), wenngleich die damalige Ausgangssituation und Zielsetzung aus heutiger Sicht von geringerer Bedeutung sind. Deutlich zunehmende wissenschaftliche Aktivitäten im Rahmen der Small Area-Forschung, vor allem im aktuellen anwendungsbezogenen Kontext, sind seit den 1960er Jahren zu verzeichnen.

Die Einsatzgebiete dabei sind vielseitig: Beginnend bei Small Area-Statistiken, etwa im Hinblick auf Ausgleichszahlungen zwischen Kommunen und Land bzw. Staat bis hin zu zuverlässigen Schätzungen von Konsumgewohnheiten aus-

gewählter kleiner soziodemographischer Gruppen. Länder wie bspw. die USA, Kanada oder Israel arbeiten schon intensiv am Einsatz solcher Schätzmethoden. In den USA etwa kommen Small Area-Methoden bei verschiedenen Bundesprogrammen zum Einsatz: Infant and maternal health for states (NCHS), Personal income for states and counties (BEA), Post-census populations for counties (USCB) und viele andere (vgl. Lahiri, 2005, und Rao, 2003). Ein beliebtes Anwendungsgebiet von Small Area-Verfahren ist unter anderem die Schätzung von Bevölkerungszahlen als Grundlage zur Ermittlung staatlicher Transferzahlungen. In diesem Zusammenhang wurde 1980 das U.S. Bureau of the Census von mehreren Staaten und Städten der USA wegen zu gering ausgewiesener Bevölkerungszahlen, insbesondere von Minderheiten, verklagt. Die Statistiker Ericksen, Kadane und Tukey, die auch als Gutachter vor Gericht auftraten, entwickelten verbesserte Verfahren, die zuverlässige Werte für das Ausmaß der Unterschätzung ethnischer Gruppen auf regionaler Ebene lieferten. Andere konkrete Einsatzgebiete sind etwa die Ermittlung verlässlicher Daten über die Verbreitung von Drogen- und Alkoholmissbrauch in ausgewählten demographischen Gruppen oder die vom U.S. Bureau of the Census im 2-Jahres-Rhythmus durchgeführte Schätzung von Armutsdaten, auf deren Grundlage bspw. über die Verteilung finanzieller staatlicher Mittel entschieden wird (vgl. Münnich und Schmidt, 2002, S. 139 und die darin zitierte Literatur). Speziell in der Arbeit von Schaible (1996) können weitere detaillierte Informationen über den Einsatz indirekter Schätzmethoden in den USA nachgelesen werden.

In Europa befassten sich zwei international ausgerichtete Forschungsprojekte, finanziert von der Europäischen Kommission im Rahmen des 5. Forschungsrahmenprogramms, ausführlich mit der Theorie sowie mit Einsatzmöglichkeiten, begleitet durch vielseitige Simulationsrechnungen und -auswertungen. Bei diesen Projekten handelt es sich einerseits um EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs; http://www.statistics.gov.uk/methods_quality/eurarea/), koordiniert vom Office for National Statistics und andererseits um das Projekt DACSEIS (Data Quality in Complex Surveys within the New European Information Society; http://www.dacseis.de), koordiniert vom Lehrstuhl für Statistik, Ökonometrie und Unternehmensforschung der Wirtschaftswissenschaftlichen Fakultät an der Eberhard-Karls Universität Tübingen. Im Projekt EURAREA beschäftigte man sich sehr intensiv mit der Methodik einzelner Small Area-Verfahren und untersuchte deren Anwendung am Beispiel zahlreicher Simulationsrechnungen. DACSEIS, das sich im Wesentlichen auf die Untersuchung der Datenqualität in komplexen Stichprobenerhebungen konzentrierte, untersuchte ebenfalls anhand umfangreicher Simulationsstudien die Einsatzmöglichkeiten der von EURAREA betrachteten Standardmethoden am Beispiel des Deutschen Mikrozensus. Zusätzlich wurde der Einfluss von Antwortausfällen auf die Modellierungen evaluiert. Ausführliche Berichte über die Er-

gebnisse beider Forschungsprojekte können den jeweiligen Internetseiten entnommen werden.

Im Rahmen von Zensen wird zunehmend auch der Einsatz von Small Area-Verfahren diskutiert, wie etwa in der Schweiz (vgl. Renaud, 2004) und nun auch in Deutschland im Zusammenhang mit einer registergestützten Volkszählung. Die heutige Bedeutung von Small Area-Methoden in Theorie und Praxis lässt sich bspw. auch an dem aktuellen Sonderheft in Statistics in Transition erkennen (siehe http://www.stat.gov.pl/english/sit/sit73/index.htm).

2.4 Aufbau der Testgesamtheit

Aufbauend auf den Ergebnissen des Zensustests (siehe Statistisches Bundesamt, 2004) und der im Forschungsprojekt DACSEIS erzeugten Saarland-Gesamtheit (siehe Münnich und Schürle et al., 2003) wurde eine Testgesamtheit, bestehend aus Personeninformationen, wie zum Beispiel Geschlecht, Nationalität, Alter, Erwerbsstatus, Adress-, Gemeinde- und Kreiszugehörigkeit, generiert. Dabei ist anzumerken, dass der DACSEIS-Datensatz zwar auf den Daten des Mikrozensus 1996 basiert, jedoch unter Formulierung adäquater Annahmen synthetisch auf die Dimension der Grundgesamtheit verlängert wurde. In Münnich und Schürle et al. (2003) ist die Methodik der Synthetisierung ausführlich beschrieben. Auf Basis dieses synthetisierten Datensatzes wurde unter Zuhilfenahme weiterer Informationen aus dem Zensustest sowie aus der Gebäudestatistik zu Anzahl der Wohnungen und Personen pro Gebäude die Testgesamtheit konstruiert. Mit Hilfe von Zusatzinformationen aus den genannten Quellen konnten notwendige Größen wie Adressinformation, Gemeinde-, Kreis- und Regierungsbezirkszugehörigkeit, Karteileichen und Fehlbestände hinzusimuliert werden. Regierungsbezirks-, Kreis- und Gemeindekennziffern wurden innerhalb der vom Mikrozensus 1996 definierten Regionalschichten vergeben. Adressdaten entstanden in Abhängigkeit der Gebäudegrößenklassen und Auswahlbezirke. Karteileichen und Fehlbestände wurden zufällig auf Adressen verteilt, so dass die aus dem Zensustest bekannten Raten pro Gemeinde- und Adressgröße im Durchschnitt grundsätzlich eingehalten sind. Abbildung 3 zeigt eine graphische Gegenüberstellung tatsächlich vorhandener Karteileichen- und Fehlbestandsraten, ermittelt aus dem Zensustest, mit den im Datensatz realisierten Quoten.

Die zu verwendende Gesamtheit mit N=1.057.915 Individuen umfasst nun alle interessierenden Kombinationen von Personen, die korrekt registrierten Personen, die so genannten Karteileichen, also Personen, die registriert, aber physisch nicht vorhanden sind bzw. vom Interviewer nicht nachweisbar sind, sowie die Fehlbestände im Register, also Personen, die angetroffen werden, jedoch nicht im zuständigen Einwohnermeldeamt gemeldet sind. Tabelle 1 zeigt die Kodierung des Datensatzes nach Karteileichen und Fehlbeständen sowie die korrespondie-

rende Anzahl der jeweils betroffenen Einheiten. Insgesamt ist der Datensatz mit 27.384 Karteileichen und 19.237 Fehlbeständen versehen.

Tabelle 1: Karteileichen und Fehlbestände in der Testgesamtheit

Nr.	Y_{wahr}	Y_{Reg}	KL	FB	
1	1	1	0	0	registriert
2	0	1	1	0	registriert, aber nicht vorhanden
3	1	0	0	1	nicht registriert, aber vorhanden
⋮			⋮		
N		...			
$\sum$			27.384	19.237	N=1.057.915 (Saarland)

Eine Auswertung der Testgesamtheit nach Kreisen, Gemeinden – gekennzeichnet durch die verschiedenen Linien innerhalb der Kreise – und Adressgrößen kann den nachfolgenden beiden Graphen absolut und relativ entnommen werden.

In Abbildung 1 erkennt man deutlich die Heterogenität der einzelnen Gemeinden, aufgeteilt nach den sechs saarländischen Landkreisen sowie die absolute Häufigkeit einzelner Adressgrößen. In Kreis 1 hebt sich erkennbar Saarbrücken als Großstadt hervor mit einer relativ hohen Anzahl kleiner Adressen. Verglichen mit Abbildung 2 sieht man, dass auf Grund der zugrunde gelegten Informationen die jeweiligen Anteile pro Adressgröße über Gemeinden und Kreise hinweg nur geringen Schwankungen unterliegen.

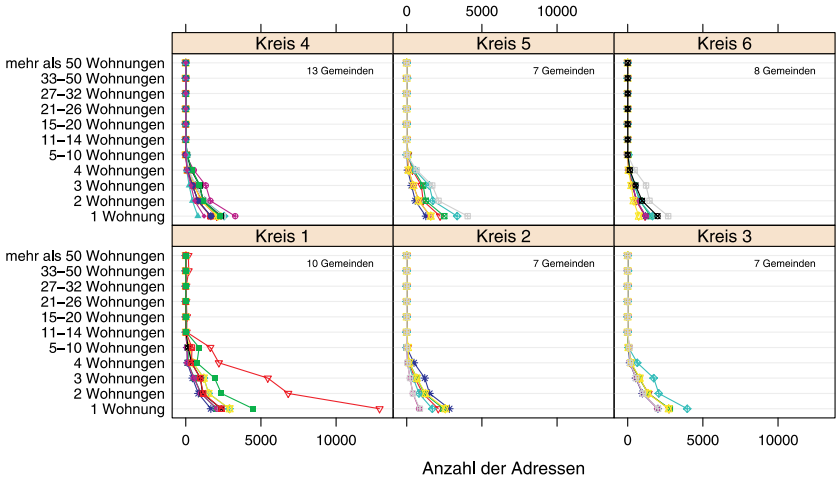


Abb. 1: Größe der Adressen in Bezug auf Kreise, Gemeinden und Adressgröße

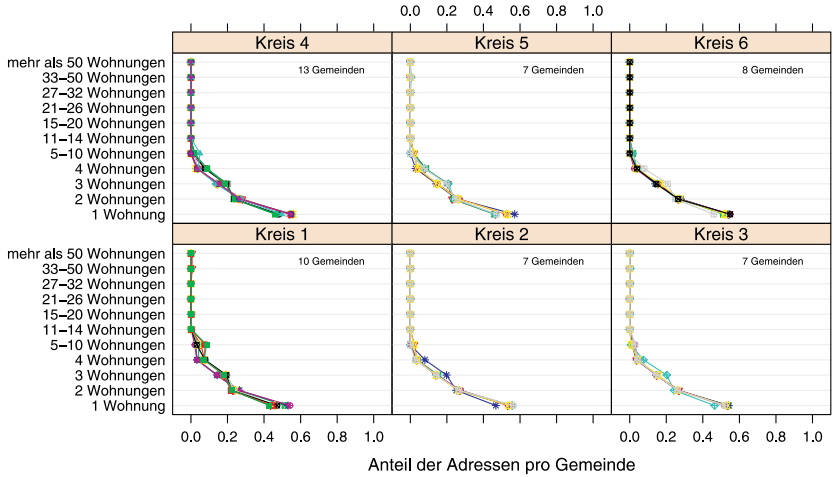


Abb. 2: Relative Größe der Adressen in Bezug auf Kreise, Gemeinden und Adressgröße

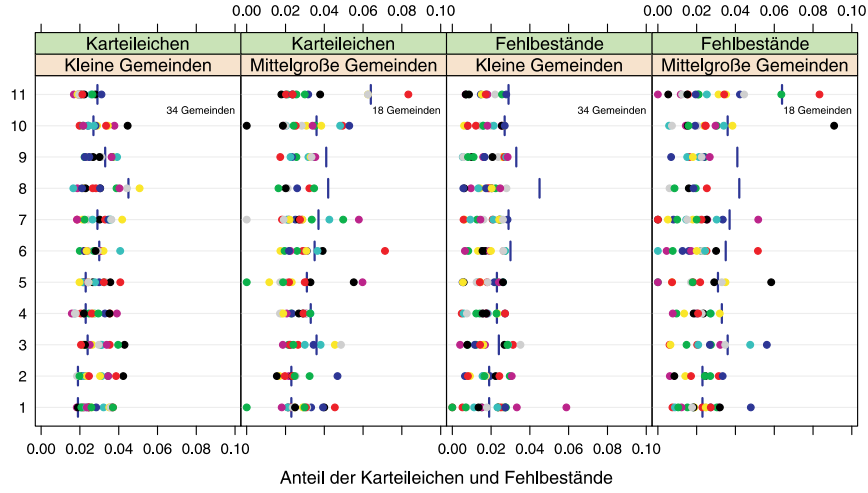


Abb. 3: Anteil der Karteileichen und Fehlbestände in Bezug auf Adress- und Gemeindegröße

In Abbildung 3 sind die Anteile der Karteileichen und Fehlbestände, aufgeteilt nach kleinen und mittelgroßen Gemeinden sowie Adressgrößen, dargestellt. Die aus dem Zensustest ermittelten Anteile sind durch die vertikale Linie und die einzelnen Gemeinden durch Punkte gekennzeichnet. Man erkennt die in der Testgesamtheit vorhandene Heterogenität der Anteile. Vereinzelt, gerade in den seltener auftretenden umfangreichen Adressgrößen, weichen die vorhandenen Anteile von Fehlbeständen und Karteileichen von den im Zensustest beobachteten Anteilen auf Grund einer unterschiedlichen Zusammensetzung der Bevölkerung geringfügig ab. Allerdings ist auch augenfällig, dass die ursprünglich erwarteten höheren Anteile von Registerfehlern bei größeren Adressen nicht so deutlich beobachtet werden können.

2.5 Interessierende Fragestellungen

Prinzipiell werden zwei grundsätzliche Fragestellungen als bevorzugt interessierend erachtet:

- Schätzung des Bevölkerungsbestandes als Basis für die Bevölkerungsfortschreibung. Diese muss eine geeignete inhaltliche sowie auch regionale Subklassifikation erlauben und von hoher statistischer Qualität sein. Als Basis dienen die Daten der Einwohnermelderegister, die um die zu schätzenden Karteileichen und Fehlbestände zu korrigieren sind.
- Schätzung von weiteren interessierenden Merkmalen auf Basis der Zusatzstichprobe. In dieser Simulation wurde die Konzentration auf die erwerbslosen Personen, die im DACSEIS-Datensatz ausgewiesen sind, gelegt.

Ausgehend vom Zensustest wurde zunächst angenommen, dass in größeren Gemeinden je 550 Adressen zufällig gezogen werden und in kleineren Gemeinden pro Kreis insgesamt 550 Adressen, die proportional im Sinne der Adressen auf die Gemeinden aufgeteilt wurden. Alternativ wurde eine uneingeschränkte Ziehung von 550 Adressen pro Gemeinde durchgeführt, wobei angemerkt werden muss, dass in der Beispielgesamtheit alle Gemeinden eine Stichprobe diesen Umfangs zuließen. Ferner wurde eine Variante des ursprünglichen Auswahlplans verwendet, bei der eine Mindestziehung von 200 Adressen realisiert wurde.

Ausgehend von der Beobachtung im Zensustest, dass bei großen Adressen eher mit Karteileichen und Fehlbeständen zu rechnen ist, wurden die drei zuvor vorgestellten einem Erhebungsdesign mit unterschiedlichen Auswahlwahrscheinlichkeiten, die proportional zur Adressgröße gewählt wurden, unterzogen. Mit diesen so genannten *probability proportional to size*-Designs (PPS) soll die Effizienz der Totalwertschätzung in Bezug auf den einfachen Horvitz-Thompson-Schätzer erheblich erhöht werden (vgl. etwa Särndal, 1992, oder Lohr, 1999).

Als Vergleichsmaßstab wurde das einfache Mikrozensus-Design gewählt (vgl. Meyer, 1994, bzw. Münnich, 2004, und die dort angegebene Literatur). Die Designs sind in nachstehender Übersicht nochmals zusammengefasst:

- Mikrozensus-Design
- Designs mit gleichen Auswahlwahrscheinlichkeiten

  GEM (550) Ziehung von je 550 Adressen pro Gemeinde

  GEM (550, prop) Ziehung von 550 Adressen in *größeren* Gemeinden ($N_i \geq 10.000$) und $N_i / N_{Kreis} \cdot 550$ in kleineren Gemeinden

  GEM (550, 200) Wie 2., aber mindestens $n_i \geq 200$
- Stichprobendesigns mit unterschiedlichen Auswahlwahrscheinlichkeiten

  GEM PPS (550) Ziehung von je 550 Adressen pro Gemeinde

  GEM PPS (550, prop) Ziehung von 550 Adressen in größeren Gemeinden ($N_i \geq 10.000$) und $N_i / N_{Kreis} \cdot 550$ in kleineren Gemeinden

  GEM PPS (550, 200) Wie 2., aber mindestens $n_i \geq 200$

Bei dieser Auswahl in Bezug auf Adressen entsteht zunächst die Frage, inwieweit die tatsächlichen Stichprobenumfänge, die bei Anwendung der verschiedenen Designs entstehen, in Bezug auf Anzahl der Haushalte und Personen und deren spezifischer Relation überhaupt miteinander vergleichbar sind. Abbildung 4 zeigt die tatsächlichen Stichprobenumfänge für Haushalte und Personen in Abhängigkeit der sieben verschiedenen Designs mit Hilfe eines Boxplots.

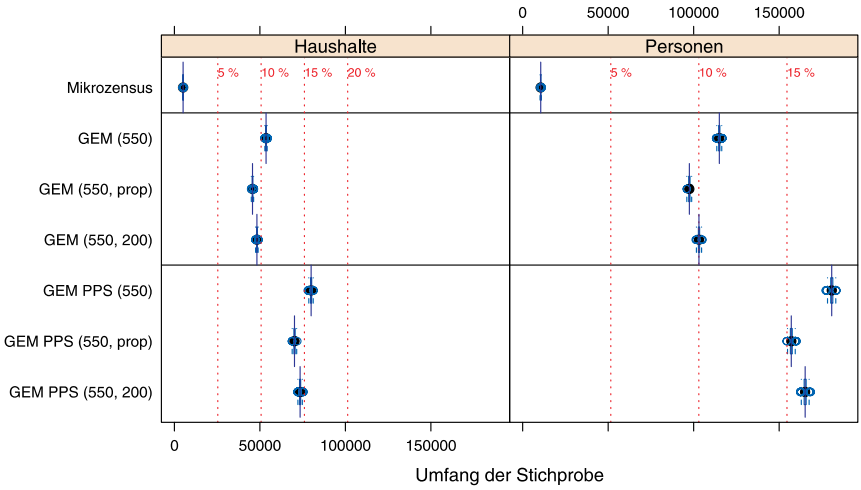


Abb. 4: Anzahl der Haushalte und Personen in den Stichproben

Der Mikrozensus eignet sich auf Grund seines geringen Stichprobenumfangs von 1% nur als Vergleichsmaßstab für die Zensus-Stichprobe, da man mit Hilfe der Zensus-Stichprobe in jedem Falle bessere Ergebnisse erzielen möchte. Die anderen Verfahren erweisen sich wie erwartet. Zunächst erkennt man, dass eine proportionale Aufteilung des Stichprobenumfangs auf kleine Gemeinden insgesamt für die geringsten Stichprobenumfänge sorgt. In Bundesländern mit vielen kleinen Gemeinden kann hier eine sehr ausgeprägte Reduktion des Gesamtstichprobenumfangs entstehen, bei der einzelne sehr ungeeignete Schätzungen zu erwarten sind, wie sich später auch zeigen wird. Eine Festlegung eines Mindeststichprobenumfangs auf n_i=200 führt hier nur zu einer mäßigen Erhöhung des Gesamtstichprobenumfangs. Auffällig erweist sich indes eine Auswahl mit unterschiedlichen Auswahlwahrscheinlichkeiten, durch welche sehr große Adresseinheiten bevorzugt gezogen werden. Hier ergibt sich nicht nur die erwartete Erhöhung des Gesamtstichprobenumfangs, sondern auch eine Verschiebung der Proportionalität von gezogenen Haushalten in Bezug auf Personen. Diese liegt darin begründet, dass im Allgemeinen bei großen Adressen nicht nur die Anzahl der Haushalte, sondern vor allem auch die Anzahl der Personen in den Haushalten überproportional anwachsen kann. Solche Disproportionalitäten sind bei einer tatsächlichen Durchführung einer Erhebung unbedingt zu beachten. Zwar scheint durch eine solche Auswahl die Effizienz der Erhebung bezogen auf die Kosten der Durchführung günstiger zu sein, die Nachteile durch eine keineswegs einfache Bestimmung von speziellen Personen- und Haushalts-spezifischen Gewichten sollten jedoch nicht unterschätzt werden. Hinzu kommt, dass möglicherweise Klumpungseffekte bei großen Adressen auftreten, die eine Schätzung zusätzlicher Variablen sehr ineffizient werden lassen kann. Dies gilt vor allem bei den möglicherweise besonders zu bevorzugenden Small Area-Verfahren (vgl. Pfeffermann et al., 1998).

3 Ergebnisse der Studie

Nachfolgend soll zunächst den zwei Hauptfragestellungen, der Bestimmung der Bevölkerungszahl sowie einer weiteren interessierenden Variablen, der Anzahl der Erwerbslosen, im klassischen Sinne nachgegangen werden. Anschließend wird noch ein Beispiel für die Small Area-Problematik gegeben.

3.1 Ermittlung des Bevölkerungsumfangs

Die ersten beiden Abbildungen zeigen Boxplots für die klassische Zensus-Fragestellung, die Ermittlung des Bevölkerungsumfangs in Bezug auf die sieben interessierenden Stichprobendesigns. Die jeweils durchgehende vertikale Linie kennzeichnet den wahren Bevölkerungsumfang in der Simulation. Auf Grund der

sehr hohen Variabilität des Horvitz-Thompson-Schätzers wurde die Skalierung der Boxplots für die effizienteren Designs optimiert.

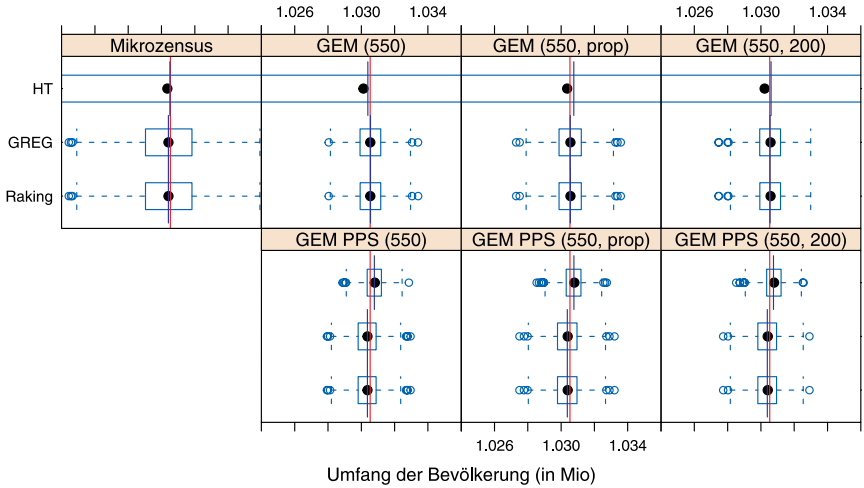


Abb. 5: Bevölkerungsumfang (Register)

In Abbildung 5 wurden beim GREG- und Raking-Schätzer als Hilfsinformation die Anzahl registrierter Personen verwendet. In der nachfolgenden Abbildung wurde zusätzlich noch die Adressgröße als Hilfsvariable hinzugezogen. Zunächst erkennt man in Abbildung 5, dass die auf Kalibrierung basierenden Schätzverfahren bei den Auswahlverfahren mit gleichen Auswahlwahrscheinlichkeiten stark zu bevorzugen sind. Innerhalb dieser Verfahren ergeben sich nur geringfügige Unterschiede, wobei sich erwartungsgemäß die kleineren Stichprobenumfänge bei kleinen Gemeinden negativ auf die Schätzungen auswirken.

Die Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten führen tatsächlich noch einmal zu einer Reduktion der Variabilität der Schätzungen. Allerdings muss hier davon ausgegangen werden, dass diese weitgehend auf die höheren Stichprobenumfänge zurückzuführen ist. Besonders auffällig ist hier die Effizienz des Horvitz-Thompson-Schätzers, die auf die Verwendung des *probability proportional to size*-Designs mit hochkorrelierter Hilfsvariablen zurückzuführen ist.

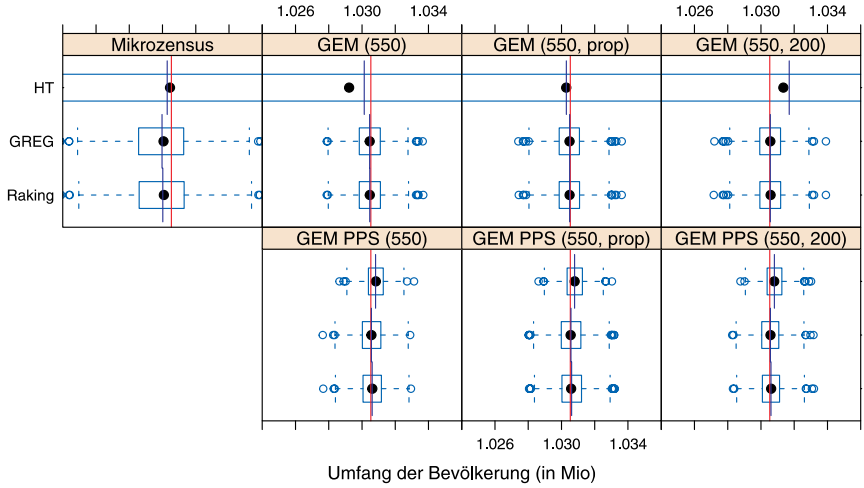


Abb. 6: Bevölkerungsumfang (Register, Adressgröße)

Die Verwendung der zusätzlichen Hilfsvariablen Adressgröße zeigt in Abbildung 6 eine geringfügige Reduktion einer möglichen Verzerrung bei den Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten. Allerdings muss hier betont werden, dass die Anzahl der Monte-Carlo-Wiederholungen mit R=1.000 nicht allzu groß ist, und die Unterschiede zum Teil unterhalb der gegebenen Monte-Carlo-Genauigkeit liegen. Abschließend sei angemerkt, dass alle Zensus-Designs auf Grund der erheblich höheren Stichprobenumfänge deutlich bessere Ergebnisse liefern als das korrespondierende Mikrozensus-Design.

3.2 Ermittlung der Anzahl der Erwerbslosen

Während im vorangegangen Fall der Bevölkerungsschätzung mit Hilfe von Registerdaten eine zur Untersuchungsvariablen sehr hoch korrelierte Variable in Gestalt der Registervariablen zur Verfügung steht, muss bei der Schätzung von Erwerbslosen zum Teil auf eine so hochkorrelierte Variable verzichtet werden. Die drei nachfolgenden Fälle unterscheiden sich wiederum in der Verwendung der Hilfsvariablen.

Eine besondere Beachtung gilt hier einer möglichen Verwendung der Daten der Bundesagentur für Arbeit. Auch wenn ein exaktes Matching zwischen den Nürnberger Registerdaten und den Daten im Zensus technisch eventuell problematisch bzw. aus Gründen des Datenschutzes nicht angebracht ist, kann in Analogie zum Mikrozensus durch Einführung einer Variablen über die Meldung als Arbeitsloser bei gleichzeitiger Übermittlung von Arbeitslosenzahlen in Gemeinden

eine Kalibrierung durchgeführt werden. Eine mögliche Auswirkung von falschen Angaben wurde in Wiegert und Münnich (2005) untersucht.

Als Hilfsvariablen wurden zunächst die in der Bevölkerungsfortschreibung vorhandenen Variablen männlich / weiblich (M/F) sowie deutsch / nicht deutsch (D/A) in Kombination verwendet; auf eine der vier Kombinationen muss jedoch aus Gründen der Kollinearität verzichtet werden. Des Weiteren wurde die Hilfsvariable arbeitslos (ALO) sowie die Kombination beider Fälle betrachtet.

Man erkennt im Vergleich zur Bevölkerungsauszählung, dass in Abbildung 7 die Hilfsvariablen keinen nennenswerten Effizienzgewinn ermöglichen. Ebenso scheinen die Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten zu kleinen Verzerrungen zu führen. Hier scheint ein Auswahlverfahren mit gleichen Auswahlwahrscheinlichkeiten bei der GREG-Schätzung eine robuste Alternative zu sein.

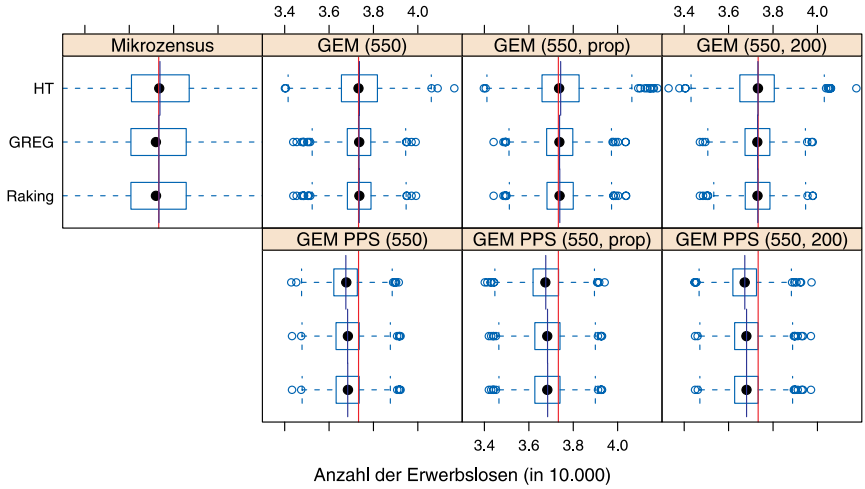


Abb. 7: Anzahl der Erwerbslosen (DM, DF, AM)

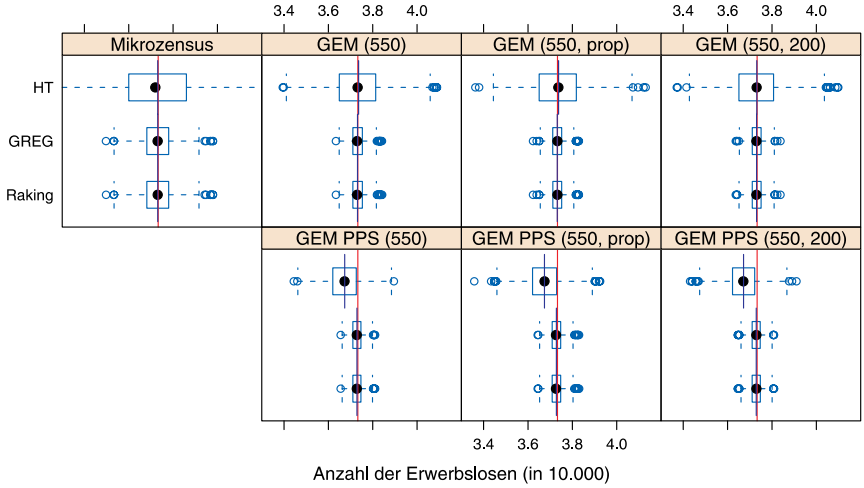


Abb. 8: Anzahl der Erwerbslosen (ALO)

Auf Grund der Tatsache, dass die Variable ALO hoch korreliert zur interessierenden Variablen der Erwerbslosen ist, überrascht es kaum, dass in Abbildung 8 wiederum ein erheblicher Effizienzgewinn zu beobachten ist. Man erkennt ebenso, dass die Verwendung der Kombination aller Hilfsvariablen keinen erkennbaren weiteren Effizienzgewinn erlaubt.

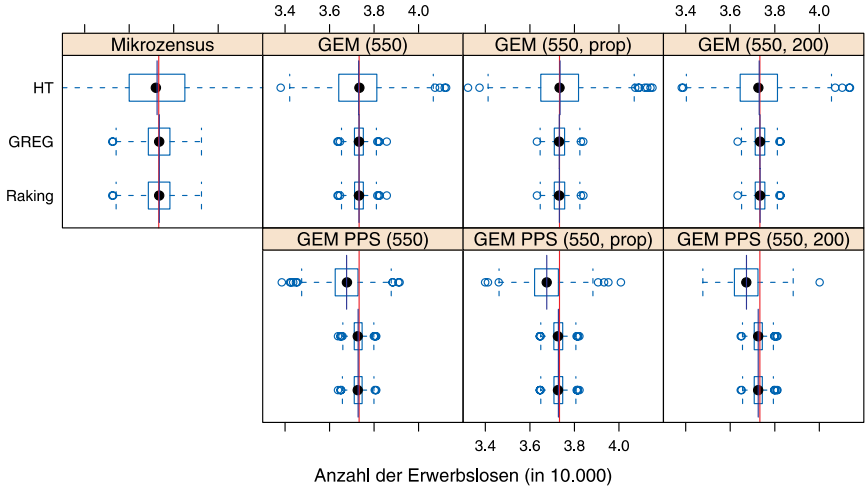


Abb. 9: Anzahl der Erwerbslosen (ALO, DM, DF, AM)

3.3 Design-Effekte

Bei beiden zuvor gestellten Fragestellungen interessiert der mögliche Effizienzgewinn bzw. -verlust des gewählten Stichprobendesigns in Bezug auf den einfachen Fall der uneingeschränkten Zufallsstichprobe. Dabei wird unter dem Design-Effekt der Quotient aus Varianz der Schätzfunktion bei gegebenem Design und bei uneingeschränkter Zufallsstichprobe und identischem Stichprobenumfang verstanden (vgl. bspw. Gabler et al., 2003, bzw. Münnich, 2005, S. 41 ff.). Es sei angemerkt, dass hier einige leicht differierende Definitionen in der Literatur existieren.

Ein Design-Effekt oberhalb von eins bedeutet demnach einen Effizienzverlust und unterhalb von eins einen Effizienzgewinn. Es wird hier jedoch nur die Wirkung des Stichprobendesigns auf einen gegebenen – vergleichbaren – Schätzer untersucht. In der Praxis muss jedoch ein Design-Effekt ebenso aus der Stichprobe geschätzt werden, was vielfach auf Grund nicht vorhandener Informationen sehr problematisch sein kann.

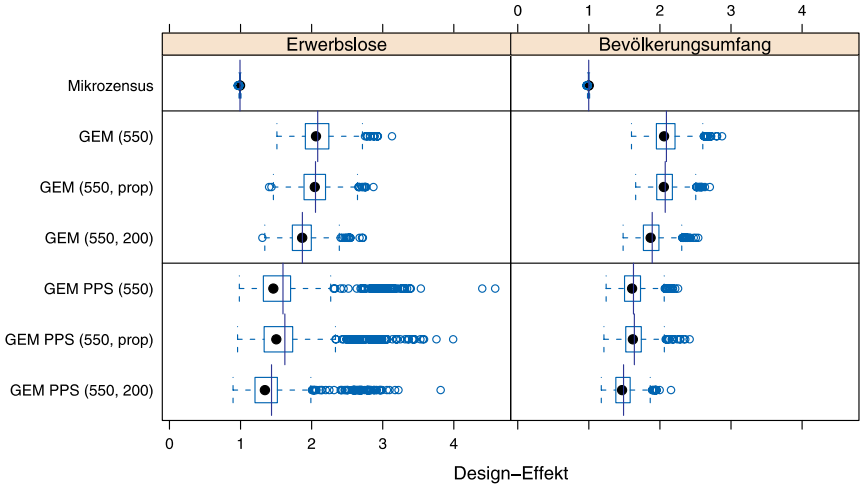


Abb. 10: Geschätzte Design-Effekte beim GREG

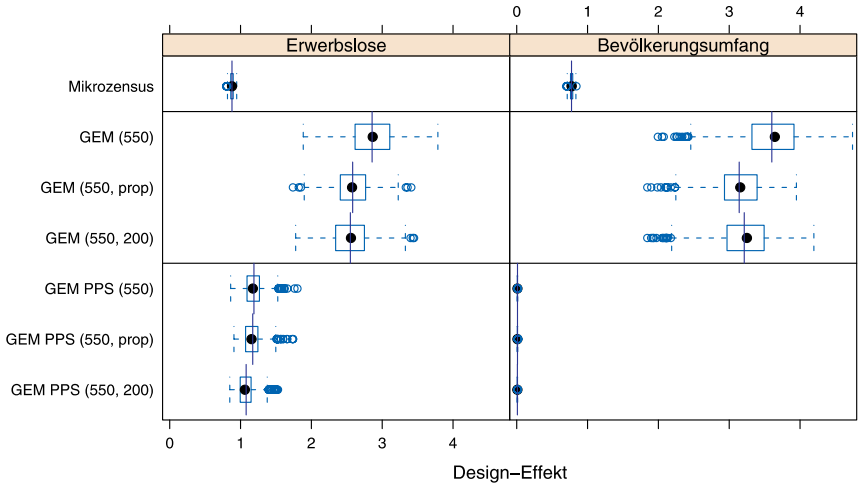


Abb. 11: Geschätzte Design-Effekte beim HT

Man erkennt, dass bei Verwendung von Adressen mit gleichen Auswahlwahrscheinlichkeiten vor allem beim Horvitz-Thompson-Schätzer auffällige Design-Effekte resultieren. Beim Mikrozensus erhält man oft Design-Effekte um eins (siehe auch Münnich, 2005, S. 154 ff.). Ebenso auffällig gestaltet sich die Verwendung von Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten. Insbesondere beim Horvitz-Thompson-Schätzer erkennt man das effiziente Design bei der Schätzung der Bevölkerung. Allerdings darf man die Verfahren untereinander nicht ohne weiteres vergleichen.

3.4 Small Area-Fragestellungen

Nach der Untersuchung von Totalwerten entsteht nun die Frage, inwieweit sich die vorher beobachteten Ergebnisse auf Gemeindeebene übertragen lassen. In den Abbildungen 12 und 13 sind die Horvitz-Thompson-Schätzwerte der Bevölkerungsumfänge absolut und relativ in den 52 Gemeinden in Bezug auf die sieben Designs mit Hilfe von Boxplots dargestellt. Man erkennt, dass die Variabilität der Ergebnisse insbesondere bei relativen Größen auffällig ist. Während bei Auswahlverfahren mit gleichen Auswahlwahrscheinlichkeiten recht große Unterschiede zwischen den Schätzungen auftreten – wie bereits bei den Gesamtwerten festgestellt –, eignen sich hier wiederum die Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten. Die proportionale Aufteilung von 550 Adressen auf kleine Gemeinden, hier mit GEM 2 gekennzeichnet, zeigt unmittelbar die kleineren Gemeinden in Form erheblich stärker variierender Schätzwerte.

Die Mindeststichprobenumfänge von n_i=200 führen unmittelbar zu stabileren Schätzergebnissen.

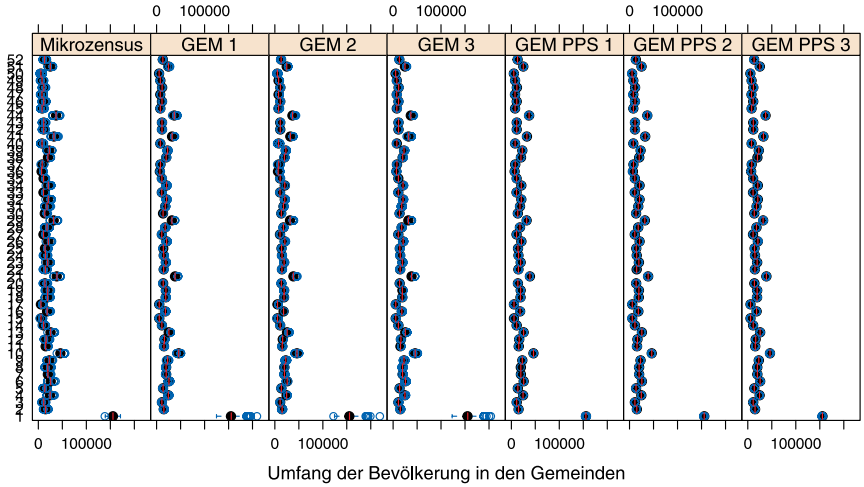


Abb. 12: Bevölkerungsumfänge in den Gemeinden (HT)

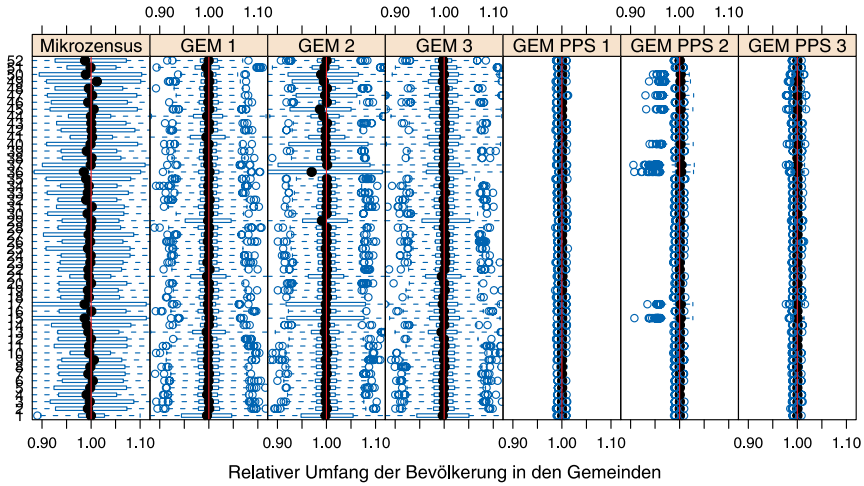


Abb. 13: Relative Bevölkerungsumfänge in den Gemeinden (HT)

Beim Regressionsschätzer ergibt sich das gleiche Bild. Allerdings scheinen hier Auswahlverfahren mit gleichen Auswahlwahrscheinlichkeiten die geringfügig

stabileren Ergebnisse zu liefern. Es sei angemerkt, dass hier ein Regressionsmodell kombiniert für alle Regionen geschätzt wurde und damit auch in die 52 Schätzergebnisse eingeht.

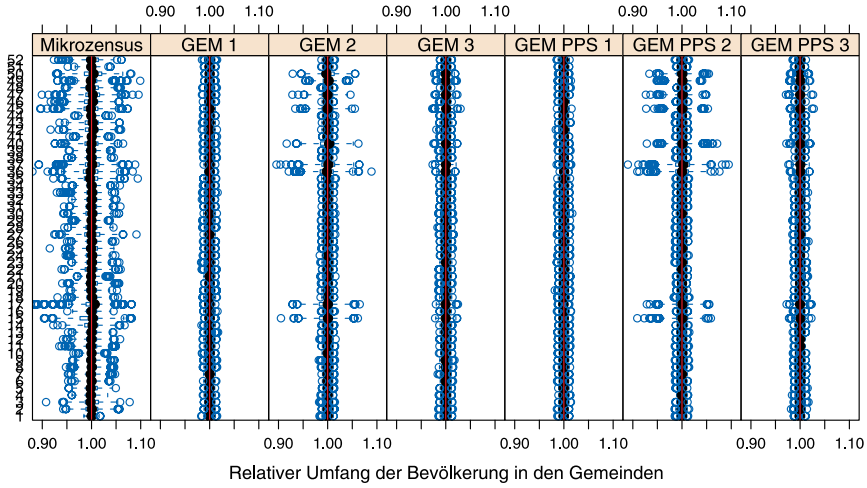


Abb. 14: Relative Bevölkerungsumfänge in den Gemeinden (GREG, kombiniert)

Abschließend wird noch ein Vergleich von Horvitz-Thompson-Schätzer, GREG und dem eigentlichen Small Area-Schätzer nach dem Modell A im Vergleich zwischen Auswahlverfahren mit gleichen und ungleichen (mit P gekennzeichnet) Auswahlwahrscheinlichkeiten dargestellt. In allen Gemeinden wurden 550 Adressen gezogen. Man erkennt wiederum eine geringere Abhängigkeit des GREG-Schätzers vom Auswahlverfahren, auch wenn beim Auswahlverfahren mit unterschiedlichen Auswahlwahrscheinlichkeiten geringfügige Verzerrungen auftreten. Der Horvitz-Thompson-Schätzer benötigt wiederum das *probability proportional to size*-Design, um überhaupt in Betracht gezogen werden zu können. Der synthetische Schätzer zeigt nun genau die bekannten Vor- und Nachteile. Im Idealfall kann man eine erhebliche Varianzreduktion erreichen. Die Modellierung zwischen den Gemeinden, also den Small Areas, ist jedoch entscheidend für das Ausmaß von Verzerrungen der Schätzungen einzelner Gemeinden.

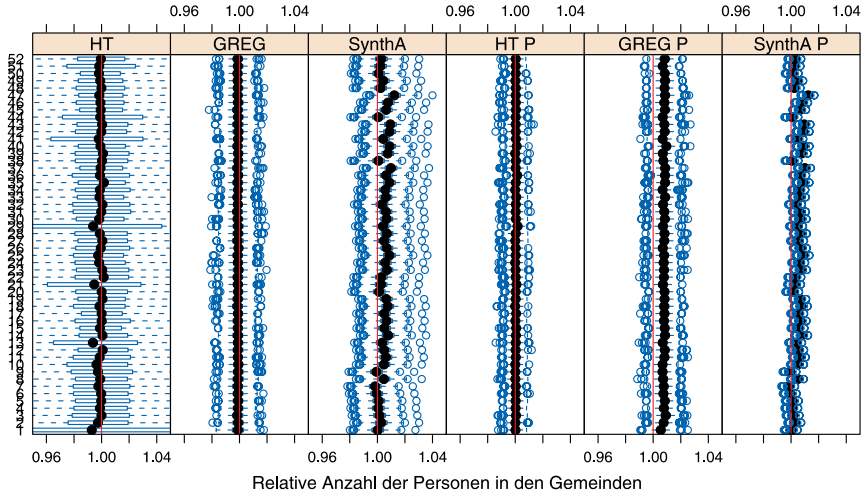


Abb. 15: Relative Bevölkerungsumfänge in den Gemeinden (550 / PPS 550)

Tiefer gehende Modellierungen erlauben hier sicherlich weitergehende Verbesserungen und damit auch eine Reduktion der beobachtbaren Verzerrung. Ebenso muss hinzugefügt werden, dass sich eine Zunahme der Anzahl der Small Areas sehr positiv auf solche Modellierungen auswirkt, da Unterschiede zwischen den Small Areas viel differenzierter geschätzt werden können. Übertragen auf die Problematik im Zensus heißt das natürlich, dass durchaus über eine Bundesländer übergreifende Verfahrensweise nachgedacht werden sollte, bei der bestimmte Klassifikationen der Gemeinden gemeinsam behandelt werden sollten.

3.5 Probleme und Einschränkungen

Die vorliegenden Untersuchungen basieren auf einer praxisnahen, aber synthetischen Gesamtheit des Saarlandes. Auf Grund der speziellen Struktur dieser Gesamtheit in Bezug auf Kreise und Gemeinden sowie Personen und Variablen muss man sowohl bei einer konkreten Übertragung auf das Saarland sowie auch bei einer Verallgemeinerung der Ergebnisse gewisse Einschränkungen in Kauf nehmen. Beispielsweise ist Rheinland-Pfalz durch eine Vielzahl kleinerer Gemeinden und vor allem Berlin, durch seine Heterogenität zwischen den Stadtteilen, in der Struktur der Variablen sicherlich anders zu bewerten. Im Falle von Berlin macht sogar die Betrachtung der Stichprobendesigns auf Gemeindeebene kaum Sinn, da Berlin eine einzige Gemeinde darstellt. Vielmehr sollten hier kleinere Einheiten herangezogen werden. Dennoch zeigen bereits diese Ergebnisse, welches Potential in modellbasierten Schätzungen liegen kann und wie stark das

Stichprobendesign die Ergebnisse beeinflussen kann. Dabei erweist sich eine konkrete Optimierung eines Sachverhalts, etwa der Schätzung des Bevölkerungsumfangs als besonders bedeutende Aufgabe der amtlichen Statistik – wie eigentlich erwartet, – als nicht unproblematisch.

Bei der Anwendung von modellbasierten Verfahren sollte aber nicht vergessen werden, dass ein hinreichend gutes Methodenwissen sowie ein Gefühl für die Daten unabdingbar ist, da nur das Zusammenspiel zu einer wirklich effizienten Schätzung führen kann. Dabei muss beachtet werden, dass geeignete Algorithmen, die im Rahmen dieser Datenmengen eingesetzt werden können, auch tatsächlich implementiert werden. Noch nicht hinreichend untersucht ist die Frage, inwieweit hier statt der klassischen Algorithmen evtl. Bayes-Methoden vorzuziehen sind.

4 Zusammenfassung und Ausblick

Diese erste Simulationsstudie hat trotz ihrer eingeschränkten Aussagekraft, bedingt durch den Umfang der Beispielgesamtheit, gezeigt, dass für die Findung eines geeigneten Designs ganz besonders vorsichtig vorgegangen werden muss. Eine mögliche Optimierung des Designs in Bezug auf die Bevölkerungszahl kann eine Reduktion der Schätzqualität der auf Small Area-Methoden basierten Schätzungen von zusätzlichen Variablen und umgekehrt bedeuten. Allerdings erscheint eine geeignete Wahl eines robusten Designs, das eine ungeeignete Klumpenbildung vermeidet, für eine Vielzahl von Fragestellungen sinnvolle Schätzergebnisse zu ermöglichen.

Ebenso darf bei aller Euphorie über die Methoden und deren Effizienz nicht vergessen werden, dass die Ziele eines Zensus klar abgegrenzt werden müssen. Dies ist nicht Aufgabe des Methodikers, sondern unabdingbare Grundlage für eine geeignete Optimierung der Verfahren in Bezug auf die möglicherweise konkurrierenden Ziele im Zensus 2011. Dabei müssen auch die Bewertungsmaßstäbe geeignet festgelegt werden. Neben der klassischen Problematik, Reduktion der Verzerrung der Schätzung versus Effizienzgewinn, der ausgerechnet beim Vergleich von klassischen und modellbasierten Methoden eine besondere Rolle zukommt (vgl. etwa Münnich et al. 2003 und Münnich et al. 2004), müssen auch die inhaltlichen Ziele auf entscheidungstheoretischer Grundlage hin quantifizierbar sein.

Es bleibt anzumerken, dass einer geeigneten Nutzung von Registerdaten sicher die Zukunft gehört. Gegenüber einer klassischen Totalerhebung, bei der durchaus auch erhebliche Messfehler auftreten können, wie etwa durch Antwortausfälle, fehlerhafte Antworten und Aufbereitungsfehler, muss jedoch die ein oder andere Einschränkung in Kauf genommen werden, insbesondere bei der Auswertung

hoch differenzierter Variablen in sehr kleinräumigen Gebieten. Insofern bedeutet der registergestützte Zensus einen Paradigmenwechsel in der deutschen Statistik, der eine Herausforderung für die Statistik-Nutzer wie auch die Methodiker sein wird. In Zeiten knapper öffentlicher Mittel ist ein klassischer Zensus aber kaum noch vertretbar. Ebenso sollte man – nach erfolgreicher Verwendung von Registern bei der Schätzung – die Vorteile solcher Schätzverfahren nicht für gering erachten.

Danksagung

Die Autoren danken Herrn Josef Schäfer, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, für die große Unterstützung bei der Generierung der Simulationsgesamtheit sowie Herrn Dr. Rolf Wiegert, Universität Tübingen, für zahlreiche wertvolle Hinweise.

5 Literaturangaben

D'Arrigo, J. und Skinner, C. J. (2003): Variance estimation for estimators subject to raking adjustment. DACSEIS deliverable D8.1, http://www.dacseis.de.

Davison, A. C.; Münnich, R.; Skinner, C. J.; Knottnerus, P. und Ollila, P. (2004): The DACSEIS Recommended Practice Manual. DACSEIS deliverable D12.3, http://www.dacseis.de.

Deville, J. C. (1999): Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. In: Survey Methodology, 25 (2), S. 193–203.

Eppmann, H. (2004): Von der Volkszählung 1987 zum registergestützten Zensus 2010? In: Statistische Analysen und Studien, 17, S. 3–9.

Gabler, S., Häder, S. und Lynn, P. (2003): Refining the concept and measurement of design effects. In: Institue, I. S. (Hrsg.): Bulletin of the International Statistical Institute 54th Session, Contributed Papers, Band LX, Book 3, S. 371–372.

Ghosh, M. und Rao, J.N.K. (1994): Small Area Estimation: An Appraisal. In: Statistical Science, 9, S. 55 - 93.

Goldstein, H. (1995): Multilevel statistical models. Second edition. London: Arnold.

Horvitz, D. G. und Thompson, D. J. (1952): A Generalization of Sampling Without Replacement From a Finite Universe. In: Journal of the American Statistical Association, 47, S. 663–685.

Lahiri, P. (2005): Workshop 15. - 16. März 2005: Introduction to Small Area Estimation. ZUMA Mannheim 2005. Unveröffentlichtes Manuskript.

Lohr, S. (1999): Sampling: Design and analysis. Pacific Grove, CA: Duxbury Press.

Longford, N.T. (1993): Random Coefficient Models. Oxford: Oxford University Press. (Oxford statistical science series; 11).

Longford, N.T. (2005): Missing data and small-area estimation: modern analytical equipment for the survey statistician. Heidelberg: Springer. (Statistics for social science and public policy).

Magg, K., Münnich, R. und Schäfer, J. (2006): Small Area Estimation beim Zensus 2011. In: Forschungsdatenzentrum der Statistischen Landesämter (Hrsg.): Amtliche Mikrodaten für die Sozial- und Wirtschaftswissenschaften. Beiträge zu den Nutzerkonferenzen des FDZ der Statistischen Landesämter 2005. (Im Erscheinen).

Meyer, K. (1994): Zum Auswahlplan des Mikrozensus ab 1990. In: Gabler, S., Hoffmeyer-Zlotnik, J. und Krebs, D. (Hrsg.): Gewichtung in der Umfragepraxis, Opladen: Westdt. Verlag.

Münnich, R. (2004): Varianzschätzung im Deutschen Mikrozensus und dessen Bedeutung für die Qualität der Angabewerte. In: Statistische Analysen, Kolloquium 2002 in Baden-Württemberg (2/2003), S. 15–23.

Münnich, R. (2005): Datenqualität in komplexen Stichprobenerhebungen. Universität Tübingen, unveröffentlichte Habilitationsschrift.

Münnich, R. et al. (2003): Data Quality in Complex Surveys. DACSEIS deliverable D1.1, http://www.dacseis.de.

Münnich, R.; Magg, K.; Sõstra, K.; Schmidt, K. und Wiegert, R. (2004): Variance Estimation for Small Area Estimates. DACSEIS deliverables D10.1 and D10.2, http://www.dacseis.de.

Münnich, R. und Schmidt, K. (2002): Small Area Estimation in der Bevölkerungsstatistik. In: Baden-Württemberg in Wort und Zahl, 3/2002, S. 139 - 145.

Münnich, R. und Schürle, J. et al. (2003): Monte-Carlo Simulation Study of European Surveys. DACSEIS deliverables D3.1 and D3.2, http://www.dacseis.de.

Pfeffermann, D.; Skinner, C. J.; Holmes, D. J.; Goldstein, H. und Rasbash, J. (1998): Weighting for Unequal Selection Probabilities in Multilevel Models. In: Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 60, No. 1, S. 23-40.

Rao, J.N.K. (2003): Small Area Estimation. Hoboken: John Wiley & Sons. (Wiley series in survey methodology).

Renaud, A. (2004): Swiss Federal Statistical Office (Hrsg.): Methodology Report: Coverage Estimation for the Swiss Population Census 2000: Estimation Methodology and Results. Neuchâtel: Swiss Federal Statistical Office. (Series: Swiss Statistics).

Särndal, C.-E., Swensson, B. und Wretman, J. (1992): Model Assisted Survey Sampling. New York: Springer.

Schäfer, J. (2004): Ergänzende Verfahren für einen künftigen registergestützten Zensus. In: Statistische Analysen und Studien, 17, S. 20–27.

Schaible, W.L. (1996) (Hrsg.): Indirect Estimators in U.S. Federal Programs. Heidelberg: Springer. (Lecture Notes in Statistics; Vol. 108).

Statistisches Bundesamt (2004) (Hrsg.): Ergebnisse des Zensustests. (Wirtschaft und Statistik; 8/2004), S. 813 - 833.

Konsequenzen der Panelmortalität im SOEP für Schätzungen der Lebenserwartung

Rainer Schnell, Mark Trappmann

1 Einleitung

Aufgrund der zahlreichen Restriktionen der amtlichen Statistik dient das Sozioökonomische Panel (SOEP) (SOEP Group 2001) auch als Datengrundlage für empirische Untersuchungen zu Gesundheit und Lebenserwartung in der Bevölkerung der Bundesrepublik. Aufsehen erregende Ergebnisse persistierender sozialer Ungleichheit bezüglich der Lebenserwartung (z.B. Klein/Unger 2001, Unger 2003) basieren daher auf den Daten des SOEP. Um Artefakte in Hinsicht auf vermeintliche soziale Ungleichheiten der Lebenserwartung zu vermeiden, muss bei solchen Analysen ausgeschlossen werden können, dass eventuell beobachtete Unterschiede in der Lebenserwartung durch unterschiedliche Austrittswahrscheinlichkeiten aus dem Panel bedingt sind.

2 Ausfallmechanismen

Bei Panelstudien besteht allgemein die Gefahr, dass Analysen durch den Austritt ehemaliger Teilnehmer aus dem Panel verzerrt sind. Nutzt man Daten aus Panelstudien für Ereignisanalysen, so bezeichnet man Beobachtungen von Untersuchungseinheiten als „zensiert", wenn das untersuchte Ereignis (z.B. Tod, Scheidung, Arbeitsplatzwechsel) bei diesen Untersuchungseinheiten nicht während des Beobachtungszeitraums eintritt. Die Ursachen hierfür können vielfältig sein: Der Beobachtungszeitraum kann zu kurz sein, das Ereignis kann bei einigen Untersuchungseinheiten niemals eintreten oder Untersuchungseinheiten können vor dem Eintritt des Ereignisses aus dem Panel austreten. Nutzt man ein Panel mit Ausfällen für Längsschnittanalysen, so können Verzerrungen vor allem dann auftreten, wenn der Grund für die Zensierung mit den zu untersuchenden Merkmalen zusammenhängt. Dies gilt besonders dann, wenn die Tatsache der Zensierung nicht durch weitere Variablen in der Studie erklärt werden kann. In einem solchen Fall bezeichnet man die Zensierung der Daten als „informativ".[1] Alle üblichen

1 Eine Zensierung kann sogar dann informativ sein, wenn keine Austritte aus dem Panel vorliegen und alle Beobachtungen zum selben Zeitpunkt enden. Das hängt damit zusammen, dass nicht alle Untersuchungseinheiten zu diesem Zeitpunkt die gleiche Ereigniszeit besitzen

Modelle zur Ereignisanalyse setzen grundsätzlich voraus, dass die Zensierung nicht informativ ist (vgl. z.B. Singer/Willett 2003, 319).

Bei Gesundheitsvariablen wie z.B. der Lebenserwartung liegt die Vermutung nahe, dass Zensierungen informativ sein könnten. So ist es plausibel anzunehmen, dass Personen mit schlechtem Gesundheitszustand ceteris paribus eher die Teilnahme verweigern als gesunde Personen. Entsprechend werden Personen mit höherer Wahrscheinlichkeit ein Interview verweigern, wenn sich ihr Gesundheitszustand seit der vorherigen Welle (an der sie noch teilgenommen haben) verschlechtert hat. Trifft diese Annahme zu, so ist die Zensierung informativ. In einem solchen Fall können die resultierenden Verzerrungen weder durch Gewichtungen noch durch Imputationen korrigiert werden.

3 Konzeptualisierung von Ausfällen und Gewichten im SOEP

Die SOEP-Arbeitsgruppe im DIW verwendet zur Korrektur der Ausfälle ein Gewichtungsmodell, bei dem zunächst zwischen Kontaktverlust und Verweigerung unterschieden wird.[2] Die Wahrscheinlichkeit für diese beiden Ereignisse wird mithilfe je eines Logit-Modells vorhergesagt.[3] Der Kehrwert des Produkts dieser beiden vorhergesagten Wahrscheinlichkeiten wird zur Längsschnittgewichtung verwendet. Allgemein werden solche Gewichte als „propensity weights“ bezeichnet.

4 Gesundheitsvariablen und die Gewichtung des SOEP

Hängt die weitere Teilnahme nur von beobachteten und dazu auch noch im jeweiligen Regressionsmodell verwendeten Merkmalen ab, dann kann ein solches Gewichtungsverfahren Verzerrungen durch Ausfälle vollständig korrigieren. Im SOEP wird dieses Ziel in Hinsicht auf gesundheitsbedingte Ausfälle nicht erreicht: Heller und Schnell (2000) konnten nachweisen, dass andere, zusätzlich zu den im Propensity-Weighting-Modell verwendeten, Variablen einen deutlichen

müssen. So gibt es unter den Befragten der ersten Welle des SOEP im Jahre 2004 beispielsweise Personen, die vom 16. bis zum 36. Lebensjahr beobachtet wurden und Personen, die zwischen ihrem 60. und 80. Lebensjahr beobachtet wurden. Die Annahme, dass die im Jahre 2004 36-jährigen zwischen ihrem 60. und 80. Lebensjahr dieselbe Sterbewahrscheinlichkeit haben werden, wie sie die dann mindestens 80-jährigen in diesen Lebensjahren hatten, ist angesichts steigender Lebenserwartung unrealistisch. Genau diese Annahme ist aber für Ereignisanalysen notwendig.

2 Vgl. z.B. Rendtel (1995), Pannenberg u.a. (2003) sowie Spieß/Kroh (2004).

3 Im Falle des SOEP werden immer nur die für den jeweiligen Wellenübergang signifikanten Regressoren aus einem Pool von Variablen verwendet, vgl. Spieß und Kroh (2004: 37ff).

Einfluss auf die weitere Teilnahme haben. Sie zeigten, dass der Tod eines Teilnehmers ein bis drei Jahre nach einer Welle auch dann noch ein guter „Prädiktor" für die Verweigerung in dieser Welle ist, wenn man alle Prädiktoren im Propensity-Weighting-Modell des SOEP und Variablen zur Messung des Gesundheitszustands kontrolliert.[4] Diese Ergebnisse weisen deutlich darauf hin, dass das Gewichtungsmodell Ausfälle durch den Zusammenhang zwischen Verweigerung und Tod allenfalls teilweise korrigieren kann.

5 Die Untersuchung der Panelausfälle: Die SOEP-Verbleibstudie

Im Jahr 2001 fand eine Verbleibstudie zum SOEP statt (Infratest Sozialforschung 2002). Mittels Adressrecherchen bei den Meldeämtern wurde der Frage nachgegangen, ob aus dem Panel ausgeschiedene Personen noch lebten und falls nicht, in welchem Jahr sie starben. 8084 Personen, deren Haushalt zwischen 1985 und 1997 durch eine Verweigerung aus dem SOEP ausschieden und die zu diesem Zeitpunkt mindestens 16 Jahre alt waren, wurden in die Verbleibstudie aufgenommen (Infratest Sozialforschung 2002:5).[5] Für 182 dieser 8084 Personen lagen keine vollständigen Informationen zum Namen vor, so dass keine Informationen bei den Meldeämtern eingeholt werden konnten. Für alle anderen konnte eine Auskunft des zuletzt zuständigen Meldeamtes eingeholt werden, wobei diese Information in 775 Fällen darin bestand, dass diese Person im Melderegister nicht mehr auffindbar ist (Infratest Sozialforschung 2002:8). Für diese Personen kann

4 Heller und Schnell verwendeten zusätzlich zu den Variablen aus dem Propensity-Weighting-Modell des SOEP Indizes für den Gesundheitszustand wie Gesundheitszufriedenheit, Krankheitsersatzindex (vgl. Fuchs und Hansmeier 1996), Pflegebedürftigkeit und Schwerbehinderung. Bei Einführung der Variable „Tod 1-3 Jahre nach der Welle" verdoppelte sich das Pseudo-r-Quadrat des Modells für beinahe alle untersuchten Wellenübergänge (Übergang a-b 0,04 auf 0,09, b-c 0,03 auf 0,11, c-d 0,04 auf 0,10, d-e 0,04 auf 0,07, e-f 0,02 auf 0,06).

5 Damit wurde nur etwa die Hälfte der bis dahin ausgeschiedenen Personen in der Verbleibstudie berücksichtigt: Bis 2001 waren 16 356 Personen aus dem SOEP ausgetreten. Ein Teil dieser Personen waren allerdings nie Befragungspersonen, sondern wurden nur in Haushaltsbögen erfasst. Neben den unter 16-jährigen und den nach 1998 Ausgeschiedenen wurden auch diejenigen nicht mit einbezogen, die durch Verschwinden des Haushalts aufgrund von Adressproblemen, Wegzug ins Ausland oder Tod ausgefallen waren, denn in „(...) all diesen Fällen liegen definitive 'Verbleib'-Informationen ja bereits vor (Infratest Sozialforschung 2002:6)". Zumindest für die aufgrund von Adressproblemen Ausgeschiedenen erscheint dies jedoch fraglich. Diese Gruppe unterscheidet sich zudem bezüglich wichtiger Variablen deutlich von den in der Verbleibstudie berücksichtigten Personen: So sind sie im Durchschnitt 8 Jahre älter (Durchschnittsalter 42,02 vs. 33,71) und stammen wesentlich häufiger (29,4 Prozent vs. 22,1 Prozent) aus der Ausländerstichprobe (Teilstichprobe B) des SOEP.

mithilfe der Verbleibstudie kein Vitalstatus festgestellt werden. Die Verbleibstudie erbrachte für 7127 Personen verwertbare Ergebnisse.

Im Verlauf der Infratest-Verbleibstudie wurden 685 Tote entdeckt, die bisher als Ausfälle mit unterschiedlichen Ursachen galten.[6] Das sind mehr als 9 Prozent der untersuchten Ausfälle. Die meisten dieser „neuen" Toten waren bereits mehrere Jahre vor ihrem Todesjahr aus dem SOEP ausgeschieden. Nur in weniger als einem Viertel der Fälle fand der letzte Kontakt im Kalenderjahr vor dem Tod oder im Todesjahr statt.[7] Im Mittel liegt das Ausscheiden mehr als fünf Jahre vor dem Tod. Aufgrund der Konzeption der Ausfälle galten diese Personen bis zur Verbleibstudie größtenteils als Verweigerer.[8] Aus den Daten der Verbleibstudie lassen sich für ein ereignisanalytisches Survival-Modell etwa 63 000 zusätzliche Episoden (Personenjahre) gewinnen.[9] Das sind 20,4% zusätzlich zu den vor der Verbleibstudie vorhandenen Episoden. Abbildung 1 fasst das Resultat der Verbleibstudie zusammen.

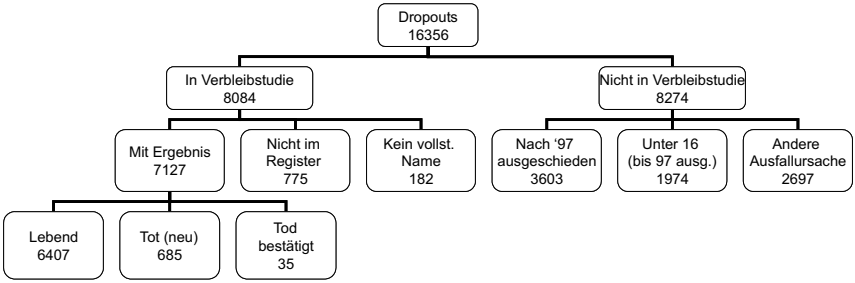


Abb. 1: Übersicht über die Ergebnisse der Verbleibstudie

6 Im Bericht von Infratest Sozialforschung ist von 720 Toten die Rede, doch handelt es sich nach Auskunft der SOEP-Arbeitsgruppe bei 35 dieser Fälle nur um eine Bestätigung einer bereits aus früheren Adressrecherchen vorliegenden Information.

7 In 25 Fällen weist das SOEP einen letzten Kontakt in einem späteren als dem Todesjahr aus (in der Regel das darauf folgende Jahr). Diese Information stammt aus der Variable „austritt" aus der Datei PPFAD, die das Jahr des letzten Kontakts angibt. Die meisten dieser Fälle stammen aus Mehrpersonenhaushalten (vor dem Tod), die (als kompletter Haushalt) im Jahr nach dem Tod die weitere Teilnahme verweigern. Der letzte Kontakt bezieht sich in diesem Fall auf den Haushalt, ist aber dennoch im Personenbearbeitungsergebnis eingetragen.

8 Für 644 der 685 neu gefundenen Toten (94,0 Prozent) lautete das letzte Personenbearbeitungsergebnis „verweigert" (Code 4 der Variable „perg").

9 Ein Personenjahr entspricht der Beobachtung einer Person in zwei aufeinander folgenden Wellen. Bezeichnet man das Eintrittsjahr einer Person i als e_i und das letzte Jahr, in dem eine Information über den Vitalstatus vorliegt (bei Toten das Todesjahr) als e_i, so gilt für die Anzahl der Personenjahre: $n = i\,(a_i - e_i)$

6 Hypothesen

Analysierte man bis zur Verbleibstudie die SOEP-Daten mit ereignisanalytischen Modellen, so traf man implizit die Annahme, dass diese Personen – gegeben ihr Alter – demselben Sterberisiko ausgesetzt sind wie gleichaltrige Teilnehmer (mit gleichen Werten bei Kovariaten). Die empirischen Befunde bei Heller/Schnell (2000) und die Analysen von Infratest (2002:11) legen hingegen nahe, dass der sich verschlechternde Gesundheitszustand in den Jahren vor dem Tod als Ursache für die Verweigerung besonders bei jenem Teil der Ausfälle plausibel ist, der wenige Jahre nach dem Ausscheiden stirbt.[10] Es ist folglich zu vermuten, dass vor der Verbleibstudie 2001 die Lebenserwartung mit dem SOEP überschätzt wurde.[11] Entsprechend sollen folgende Hypothesen hier näher untersucht werden:

1. Schätzt man die Lebenserwartung mithilfe des SOEP vor der Verbleibstudie 2001, so führt dies zu einer Überschätzung der Lebenserwartung.
2. Diese Überschätzung bleibt bestehen, wenn man die SOEP-Gewichtung verwendet.

Diese Überschätzung kann zum einen gegenüber den SOEP-Daten nach der Verbleibstudie, zum anderen gegenüber Daten der amtlichen Statistik bewertet werden. Vergleicht man die Schätzung der Lebenserwartung aufgrund des SOEP vor und nach der Verbleibstudie miteinander, so reduziert sich die Differenz auf den Ausfallmechanismus. Scheiden kranke Personen (und damit Personen mit erhöhter Sterbewahrscheinlichkeit) eher aus, so überschätzt das SOEP vor der Verbleibstudie gegenüber dem SOEP nach der Verbleibstudie die Lebenserwartung. Bestätigen sich die Hypothesen, so wäre dies ein Hinweis, dass es ohne die

10 Zur begünstigenden Wirkung von Krankheit auf Verweigerungen vgl. auch Unger (2003:73) und Dufouil/Brayne/Clayton (2004).

11 Diese Schlussfolgerung steht im Gegensatz zu den Befunden bei Rendtel (1995). Danach gab es in den ersten sieben Erhebungswellen nur geringfügige Abweichungen zwischen der aufgrund der Angaben im SOEP geschätzten Anzahl der Todesfälle und der aufgrund von Registerdaten erhobenen Zahl der Todesfälle in der Population. Laut SOEP wären 4 066 000 Todesfälle für die Population zu erwarten gewesen, stattdessen gab es 4 202 000 Todesfälle. Die Unterschätzung um ca. 3,2 Prozent ist nicht signifikant und kann zudem mit der Untererfassung der Anstaltsbevölkerung gerade in den Anfangsjahren erklärt werden (Rendtel 1995:238ff.). Allerdings war ein Versuch der Autoren, die von Rendtel berichteten Ergebnisse für dieselben Jahre mit den mit einer neueren Welle ausgelieferten SOEP-Daten zu replizieren, nicht erfolgreich. Hierfür können unter anderem nachträgliche Veränderungen der SOEP-Gewichtung ursächlich sein. Die verfehlte Signifikanz der Unterschätzung darf aber nicht als Beleg dafür interpretiert werden, dass keine Unterschätzung stattfindet. Signifikanz ist neben der Effektstärke (und der Wahl des Signifikanzniveaus) von der Fallzahl und der Varianz der verwendeten Gewichte abhängig. Da es in den von Rendtel analysierten Jahren nur 554 Tote gab und diese zudem mit sehr ungleichen Gewichten gewichtet wurden, kann die verfehlte Signifikanz auch als Ergebnis der geringen Fallzahl verstanden werden.

Durchführung von Verbleibstudien zu Verzerrungen bei der Schätzung der Lebenserwartung käme.

7 Schätzung der Lebenserwartung anhand des SOEP vor und nach der Verbleibstudie

Möchte man Aussagen über die Differenzen der Sterbewahrscheinlichkeiten in unterschiedlichen Datenquellen für dieselbe Population machen, so liegt die Verwendung einer allgemein verständlichen Maßzahl wie die Zahl der Toten nahe. Leider kann dieser einfache Weg in der Folge nicht beschritten werden. Das liegt daran, dass alte und kranke Personen im SOEP deutlich unterrepräsentiert sind[12] .

Da die Altersgruppen mit der höchsten Sterbewahrscheinlichkeit deutlich unterrepräsentiert sind, ist auch die Inzidenzrate der Sterbefälle (Anzahl der Todesfälle pro Beobachtungsjahr) in der SOEP-Stichprobe deutlich geringer als in der Population. Die Inzidenzrate des SOEP liegt vor der Verbleibstudie (ohne Gewichtung) bei nur 0,0059 (das entspricht einem Todesfall auf 169 Beobachtungsjahre). Die Inzidenzrate in der Bevölkerung liegt 2003 laut Statistischem Bundesamt dagegen bei 0,0103.[13] Das ist für sich genommen jedoch noch kein Problem und könnte prinzipiell durch das Gewichtungsverfahren des SOEP korrigiert werden. Allerdings wird das Vorgehen, die gewichtete Anzahl der Toten für das SOEP nach der Verbleibstudie zu berechnen, dadurch verhindert, dass zwar für die Jahre bis zum Austritt ein Gewicht existiert, doch existiert dieses Gewicht für die Jahre zwischen dem Ausscheiden und der Verbleibstudie nicht. Somit kann zwar für den SOEP vor der Verbleibstudie eine gewichtete Zahl der Toten hochgerechnet werden (da dann keine Episoden zwischen Austritt und Verbleibstudie zu berücksichtigen sind), für den SOEP nach der Verbleibstudie sind jedoch für die Episoden zwischen Austritt und Verbleibstudie keine Gewichte vorhanden. Eine Lösungsmöglichkeit bestünde darin, für die Jahre nach dem Ausscheiden zur SOEP-Gewichtung analoge Gewichte zu berechnen, doch ist dies problematisch: Da die Gewichte die Teilnahmewahrscheinlichkeit einbeziehen und die Teilnahme bei Einbeziehung der Daten aus der Verbleibstudie anders definiert

12 Laut Statistischem Bundesamt waren 2001 ca. 17,1 Prozent der deutschen Bevölkerung 65 Jahre und älter, im SOEP sind es nur 13,5 Prozent. 80 Jahre und älter sind etwa 4,0 Prozent in der Bevölkerung, 1,9 Prozent im SOEP.

13 Es sei darauf hingewiesen, dass diese Inzidenzrate zwar bedeutet, dass es nur einen Todesfall auf 97 gelebte Jahre gibt, dies ist aber nicht zu verwechseln mit einer durchschnittlichen Lebenserwartung von 97 Jahren, sondern hängt von der Altersverteilung in der Bevölkerung ab.

wäre (nämlich die Teilnahme an der Verbleibstudie mit einbezöge), bekämen Gruppen mit geringen Bleibewahrscheinlichkeiten daher zu hohe Gewichte.[14]

Daher wird zunächst ein ungewichtetes nicht-parametrisches ereignisanalytisches Modell ohne Kovariaten nach dem Kaplan-Meier-Verfahren (Kaplan und Meier 1958) geschätzt. Bei diesem Verfahren spielt die abweichende Altersverteilung zwischen SOEP und Bevölkerung keine Rolle, da das Modell auf der Berechnung von Hazardraten beruht, die die Wahrscheinlichkeit des Ereigniseintritts (hier: Tod) innerhalb einer Zeiteinheit, gegeben ein bestimmtes Lebensalter, ausdrücken. Technische Details der Schätzung werden in Anhang II dargestellt.

Unterscheidet man zunächst nicht zwischen unterschiedlichen Teilstichproben und Geburtskohorten, dann fällt auf, dass sich die Inzidenzrate durch Berücksichtigung der Verbleibstudie deutlich erhöht. Vor der Verbleibstudie liegt die Inzidenzrate bei 0,0059, nach der Verbleibstudie bei 0,0067. Dass die Inzidenzrate in der Verbleibstudie mit 0,0106 deutlich über der Rate im SOEP liegt, ist zu erwarten, denn es handelt sich ja um spätere Lebensjahre derselben Personen.

Das Resultat der Ereignisanalyse (siehe Tabelle 1) zeigt, dass die Ergebnisse vor und nach der Verbleibstudie sich dennoch nicht wie erwartet unterscheiden. Sämtliche Quartile der Survivor-Funktion sind für beide Modellierungen identisch. Zwar lässt sich in Tabelle 2 ablesen, dass es vor der Verbleibstudie etwas weniger Ereignisse (also: Tode) gab als bei Gleichheit der Survivor-Funktionen vor und nach der Verbleibstudie zu erwarten wäre, doch ist das Ergebnis nicht signifikant. Daran ändert sich auch nichts, wenn man die Daten nach Geschlechtern getrennt untersucht.

Tabelle 1: Survivor-Funktion vor und nach der Verbleibstudie

				survival time		
Verbl.	time at risk	incidence rate	no. of subjects	25%	50%	75%
nachher	370548	.0067414	49075	72	81	88
vorher	307660	.0059319	49034	72	81	88
Total	678208	.0063742	98109	72	81	88

14 Selbst diese fehlerhaften Gewichte können praktisch nicht ohne Abweichung gegenüber dem SOEP berechnet werden, da die Kodierungen der für das Gewichtungsverfahren des SOEP verwendeten Variablen für die älteren Wellen des SOEP nicht exakt dokumentiert wurden.

Tabelle 2: Log-Rank-Test auf Gleichheit der Survivor-Funktionen

Verbl.	Events observed	Events expected
nachher	2498	2467.81
vorher	1825	1855.19
Total	4323	4323.00

$\chi^2(1) = 0.91$

$\Pr > \chi^2 = 0.3399$

Die Konsequenzen der SOEP-Verbleibstudie sind nicht so schwerwiegend wie erwartet wurde: Survivor-Funktionen auf der Basis der SOEP-Daten vor und nach der Infratest-Verbleibstudie 2001 unterscheiden sich nicht signifikant voneinander. Die Notwendigkeit aufwändiger Verbleibstudien für Mortalitätsanalysen lässt sich daher zunächst mit den Daten der Verbleibstudie nicht belegen. Daher stellt sich die Frage, ob sich die Effekte informativer Zensierung bei relativ seltenen Ereignissen in einem Panel überhaupt nachweisen lassen. Diese Frage lässt sich am einfachsten durch eine Simulation beantworten.

8 Simulation der Konsequenzen informativer Zensierung

Für diese Simulation wird angenommen, dass die Population in zwei gleich große Teilpopulationen „H“ und „S“ zerfällt: H repräsentiert „gesunde“ und S „kranke“ Befragte. Die Mitglieder der Teilpopulation S haben gegenüber den Mitgliedern von H ein um den Faktor f erhöhtes Sterberisiko und eine geringere Bleibewahrscheinlichkeit $p_S < p_H$. Wir gehen nun vereinfachend davon aus, in der ersten Welle mit einer Stichprobe zu beginnen, in der in jedem Alter das Sterberisiko dem der Population entspricht. Da sich sowohl die Sterberisiken als auch die Bleibewahrscheinlichkeiten der beiden Gruppen unterscheiden, entsteht jedoch im Zeitverlauf eine informative Zensierung, die dazu führt, dass die aus der Stichprobe geschätzten bedingten Sterbewahrscheinlichkeiten nach Alter sich von den bedingten Wahrscheinlichkeiten in der Population unterscheiden. Mit der Simulation sollte die Frage geklärt werden, bei welchen Werten für den Faktor f und die Bleibewahrscheinlichkeiten p_S und p_H sich die aus der Stichprobe geschätzten Survivor-Funktionen signifikant von einer Survivor-Funktion unterscheiden, die aus einer unverzerrten Zufallsstichprobe derselben Größe geschätzt wurde. Technische Details der Simulation werden in Anhang I dargestellt.

Die Simulation zeigt, dass der Faktor f mindestens die Größe 1,47 haben müsste, damit die Unterschiede auf dem Niveau $\alpha = 0{,}05$ signifikant würden. Die „kranke“ Hälfte der Population müsste also bei doppelter Ausfallwahrscheinlich-

keit ein um 47 Prozent erhöhtes Sterberisiko besitzen, damit Unterschiede zwischen der Stichprobe mit informativem Ausfallprozess und der ohne Ausfälle signifikant würden. Noch anschaulicher darstellbar wird das Resultat der Simulation, wenn man die Bedingung gleicher Größen der beiden Stichproben fallen lässt, dafür aber den Faktor f festhält. Die Frage könnte dann lauten, wie groß der Anteil einer „kranken" Subpopulation mit doppeltem Sterberisiko (i.e. f = 2) sein müsste, damit sich die Stichproben mit und ohne selektiven Ausfall (erstere wieder mit $p_S = 0{,}92$ und $p_H = 0{,}96$) voneinander unterscheiden. In diesem Fall bräuchte man etwa 17 Prozent kranke Personen, um zu einem signifikanten Ergebnis beim Log-Rank-Test zu gelangen ($\alpha = 0{,}05$).

Als Fazit lässt sich aus diesen Simulationen ziehen, dass (aufgrund der relativen Seltenheit von Toden) relativ starke Zensierungsmechanismen notwendig sind, damit die informative Zensierung unter den Bedingungen des SOEP zu signifikant anderen Ergebnissen führt als eine Stichprobe ohne Austritte. Die fehlende Signifikanz der Schätzung nach der Verbleibstudie zur Schätzung vor der Verbleibstudie kann daher allein kein Argument für die Unverzerrtheit der Mortalitätsschätzung auf der Basis des SOEP darstellen.

9 Vergleich der Lebenserwartung auf der Basis des SOEP mit amtlichen Sterbetafeln

Sollten die Hypothesen über das erhöhte Ausfallrisiko von Personen mit geringerer Lebenserwartung korrekt sein, dann könnten trotz des fehlenden Effekts der Verbleibstudie die Schätzungen der Lebenserwartung auf Basis des SOEPs zu optimistisch sein. Hier sind mehrere Mechanismen denkbar:

1. Möglicherweise kann eine Gruppe mit besonders hohem Sterberisiko auch in der Verbleibstudie nicht ausfindig gemacht werden.[15]
2. Es wurden nicht alle Ausfälle bis 2001 in die Verbleibstudie einbezogen.
3. Die deutliche Untererfassung der Anstaltsbevölkerung im SOEP legt nahe, dass die Resultate sowohl vor als auch nach der Verbleibstudie die tatsächliche Sterbewahrscheinlichkeit unterschätzen.[16]

15 775 Fälle sind im Melderegister nicht auffindbar.

16 Dazu sollte man sich klarmachen, dass zur Anstaltsbevölkerung die Bewohner von Altenheimen und Hospizen zählen. Zwar werden SOEP-Befragte prinzipiell in solche Einrichtungen weiterverfolgt, doch zeigen Analysen, dass dies nicht in vollem Maße gelingt und daher die Anstaltsbevölkerung im SOEP unterrepräsentiert ist (Panneberg u.a. 2003:165: „[...] the rate of response of respondents who move to old age homes is lower than average.").

4. Schließlich ist es denkbar, dass kranke Menschen in Anbetracht der langfristig konzipierten Befragung vor allem bei der erstmaligen Rekrutierung die Teilnahme verweigerten und daher gar nicht ins SOEP gelangten.[17]

Daher werden im Folgenden beide Resultate (vor und nach der Verbleibstudie) mit offiziellen Sterbetafeln (Statistisches Bundesamt 2004) verglichen. Da für die Daten vor der Verbleibstudie Gewichte vorliegen, kann für die aus dem SOEP vor der Verbleibstudie geschätzte Survivor-Funktion eine gewichtete und eine ungewichtete Version zugrunde gelegt werden, so dass wir für Frauen und für Männer jeweils drei Versionen aus dem SOEP geschätzter Survivor-Funktionen mit den Daten der amtlichen Statistik vergleichen.

Für eine genauere Analyse können die sechs aus den SOEP-Daten geschätzten Survivor-Funktionen mit einer mit der SOEP-Fallzahl der jeweiligen Jahre gewichteten Survivor-Funktion aus amtlichen Statistiken der Jahre 1984 bis 2001 verglichen werden. Es dürften bei einer unverzerrten Stichprobe aus der Bevölkerung der Bundesrepublik lediglich Differenzen aufgrund der Stichprobenvarianz auftauchen.[18] Die Abbildungen 2 und 3 stellen die Differenz zwischen der aus dem SOEP geschätzten Survivor-Funktion (nach den drei unterschiedlichen Versionen: vorher, vorher mit Gewichtung, nachher) und der Survivor-Funktion der amtlichen Statistik dar. Abbildung 2 zeigt den Zusammenhang für Frauen. Bis zum Alter von siebzig Jahren überschätzt das SOEP den Anteil noch Lebender nach allen Versionen nur um maximal etwa einen Prozentpunkt. Danach steigt die Differenz nach allen Versionen. Erst bei einem Alter von über 80 Jahren[19] zeigt sich dann ein deutlicher Trend: Die gewichteten SOEP-Daten vorher überschätzen die Survivor-Funktion am deutlichsten, nämlich um bis zu 7,2 Prozentpunkte

17 Die Ausschöpfungsquote der ersten Welle lag bei nur knapp über 60 Prozent (Panneberg u.a. 2003:144). Vor allem in Teilstichprobe A (Deutsche (West)) war die fehlende Ausschöpfung fast ausschließlich auf Verweigerungen zurückzuführen. Besonders deutlich unterrepräsentiert waren bereits in der ersten Welle ältere Personen (Panneberg u.a. 2003:145). Dies könnte ein Hinweis auf eine erhöhte Verweigerungsneigung von Personen mit schlechtem Gesundheitszustand sein.

18 Hier mag der Einwand kommen, dies sei nicht die Grundgesamtheit des SOEP. Dieser Einwand ist nur partiell korrekt. So gehörte zwar die Anstaltsbevölkerung nicht zur Grundgesamtheit der ersten Welle, die neueren Weiterverfolgungsregeln schreiben jedoch vor, Befragte in die Anstalten weiter zu verfolgen. In späteren Wellen kann bei Gelingen der anvisierten Weiterverfolgung die Anstaltsbevölkerung nicht mehr aus der Grundgesamtheit ausgenommen werden. Zwar gelingt die Weiterverfolgung im SOEP nur teilweise, doch wird versucht, dem mit erhöhten Gewichten für institutionalisierte Personen zu begegnen (Panneberg u.a. 2003: 166). Somit gehören auch institutionalisierte Personen zur Grundgesamtheit des SOEP. Weiterhin könnten zwar prinzipiell institutionalisierte Personen aus Analysen ausgeschlossen werden, doch wären Studien zur Lebenserwartung, die die Anstaltsbevölkerung ausnähmen, inhaltlich kaum von Interesse.

19 Ab diesem Alter setzt bei Frauen verstärkt das Sterben ein: 59,0 Prozent der Frauen erreichen das 80. Lebensjahr, aber nur noch 18,9 Prozent das 90.

(während nur 18,9 Prozent der Frauen das Alter von 90 erreichen, sind es nach dem SOEP geschätzte 26,1 Prozent), während die SOEP-Daten nachher die Survivor-Funktion am wenigsten deutlich überschätzen. Die Verwendung der Gewichtung vergrößert also noch den Fehler statt ihn zu korrigieren. Alle aus dem SOEP abgeleiteten Schätzungen kommen jedoch zu einer Überschätzung der Survivor-Funktion und damit zu einer Unterschätzung der Sterbewahrscheinlichkeit.

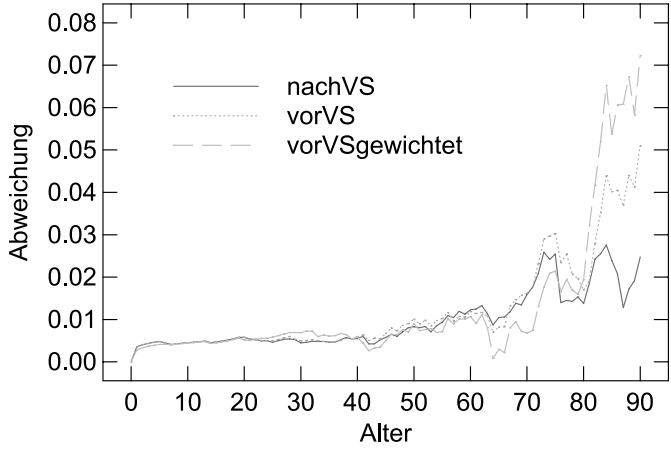


Abb. 2: Abweichungen von der Survivor-Funktion der amtlichen Statistik (weiblich)

Bei den Männern (vgl. Abbildung 3) sieht der Zusammenhang anders aus. Zwischen den ungewichteten Schätzungen aus dem SOEP vor und nach der Verbleibstudie gibt es nur geringfügige Differenzen. Beide führen zu einer Überschätzung gegenüber der amtlichen Statistik, die bis zum Alter von etwa 64 Jahren auf 4,5 Prozentpunkte ansteigt. Danach schwanken sie bei etwa 2,6-7,6 Prozentpunkten Differenz, wobei ab einem Alter von 80 die SOEP-Daten nach der Verbleibstudie durchgehend näher an der amtlichen Statistik liegen. Der Effekt der Gewichtung ist bei den Männern ein gänzlich anderer. Gewichtet man die Daten, so überschätzt man zwar auch bei den Männern immer noch die Survivor-Funktion, doch liegen die gewichteten Daten fast durchgehend näher an der amtlichen Survivor-Funktion als die ungewichteten Daten vor und nach der Verbleibstudie.

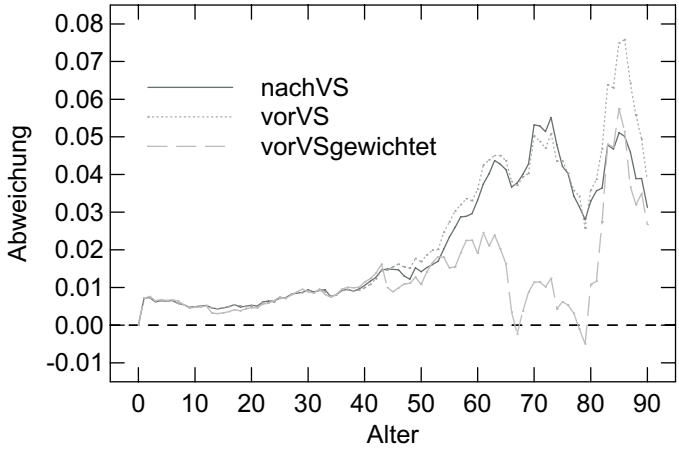


Abb. 3: Abweichungen von der Survivorfunktion der amtlichen Statistik (männlich)

Die p-Werte der Log-Rank-Tests zweiseitiger Tests auf Gleichheit gegenüber der amtlichen Statistik zeigen für beide Geschlechter überzufällige Differenzen zwischen amtlichen Daten und SOEP sowohl vor der Verbleibstudie (Frauen 0,0019, Männer: 0,0000) als auch nach der Verbleibstudie (Frauen 0,0307, Männer 0,0000).[20]

Verwendet man vor der Verbleibstudie die SOEP-Gewichtung, so zeigt sich graphisch dasselbe Bild: Das SOEP unterschätzt die Anzahl der Toten. Während die Gewichtung bei den Männern die große Differenz zwischen ungewichteter SOEP-Kurve und amtlicher Statistik etwas reduziert und damit nicht mehr signifikant ist (p=0,1026), vergrößert sich die Differenz bei den Frauen bei Berücksichtigung der Gewichtung sogar (Cox-Test: p=0,0009).

20 Um den Log-Rank-Test zum Vergleich der Survivor-Funktionen zweier Stichproben nutzen zu können, muss angenommen werden, dass die Sterbewahrscheinlichkeiten der amtlichen Statistik aus einer zweiten Stichprobe stammen und folglich mit einem Standardfehler behaftet sind. Tatsächlich handelt es sich hierbei jedoch um die Populationsparameter. Die Konsequenzen unseres Vorgehens sind in Hinsicht auf den Test konservativ, die korrekten p-Werte sind also noch kleiner. Angemessener wäre hier die Verwendung des konservativeren „one sample log-rank test“ (Finkelstein/Muzikansky/Schoenfeld 2003).

10 Mögliche Erklärung der Unterschiede zwischen dem SOEP nach der Verbleibstudie und der amtlichen Statistik

Das SOEP unterschätzt die Zahl der Toten also auch nach der Verbleibstudie. Für die noch verbleibende Unterschätzung gibt es mindestens zwei Erklärungsansätze, zwischen denen mithilfe der SOEP-Daten nicht unterschieden werden kann und die durchaus auch gemeinsam wirken können:

1. Ein Teil der Ausgeschiedenen wurde nicht in der Verbleibstudie berücksichtigt und die Verbleibstudie lieferte nicht für alle Berücksichtigten Ergebnisse. Die Personen ohne Ergebnis könnten ein gegenüber der Population erhöhtes Sterberisiko aufweisen und ihr Fehlen führte dann zu einer Unterschätzung des Sterberisikos auch nach der Verbleibstudie.
2. Nicht nur bei der Entscheidung zur Wiederteilnahme beim Wellenübergang, sondern vor allem bei der Entscheidung zur Erstteilnahme spielt der Gesundheitszustand eine Rolle. Somit könnte bereits die jeweils erste Welle einer Teilstichprobe ein zu geringes Sterberisiko gegenüber der Population aufweisen.

Die Ursachen für die erste mögliche Fehlerquelle können mit den vorliegenden Daten etwas genauer beschrieben werden. Hierzu wurde explorativ untersucht, ob sich die 775 Personen, für die die Verbleibstudie kein Resultat erbrachte, vom Rest der insgesamt 8084 Personen in der Verbleibstudie unterscheiden. Für diese Untersuchung wurde CART (*Classification and Regression Trees*, Breiman u.a. 1984) verwendet. Dabei zerlegt CART die Stichprobe durch sukzessive binäre Partitionierung in Form eines Baumes in Subgruppen, die sich bezüglich einer abhängigen Variable besonders deutlich voneinander unterscheiden.[21]

CART ermittelt als wichtigsten Prädiktor die Zugehörigkeit zu Sample B (Ausländer (West)).[22] Konnten insgesamt 10,3 Prozent der in der Verbleibstudie Berücksichtigten im Melderegister nicht mehr aufgefunden werden, so sind es 15,4 Prozent der Ausländer und nur 8,5 Prozent der Deutschen. Bei den Deutschen zeigt sich danach das Austrittsjahr als guter Prädiktor (vor 1989: 12,1 Prozent, ab 1989: 5,0 Prozent), bei den Ausländern das Geburtsjahr (vor 1926: 50,0 Prozent (n=20), ab 1926: 14,8 Prozent). Von den Befragten aus Stichprobe A

21 Die Analyse wurde mit SYSTAT berechnet. Als Zielfunktion wurde die Minimierung der Summe der quadrierten Differenzen bezüglich der abhängigen Variable innerhalb der Teilgruppen verwendet.

22 Es ist sinnvoll, auch zeitabhängige Variablen (wie z.B. die Haushaltsgröße) einzubeziehen. Daher lässt sich für jedes Jahr ein solcher Baum erstellen. Da sich aber die zeitabhängigen Variablen meist nur sehr langsam verändern, gleichen sich die Ergebnisse im Zeitverlauf. Die hier präsentierten Ergebnisse beziehen sich auf die Werte des Jahres 1984 (1. Welle). Durch fehlende Werte bei den zeitabhängigen Prädiktoren reduziert sich für diese Analyse die Zahl der Personen auf diejenigen 4399 Personen (der 5544 Personen aus der Verbleibstudie) aus der ersten Welle mit gültigen Werten für alle Prädiktoren.

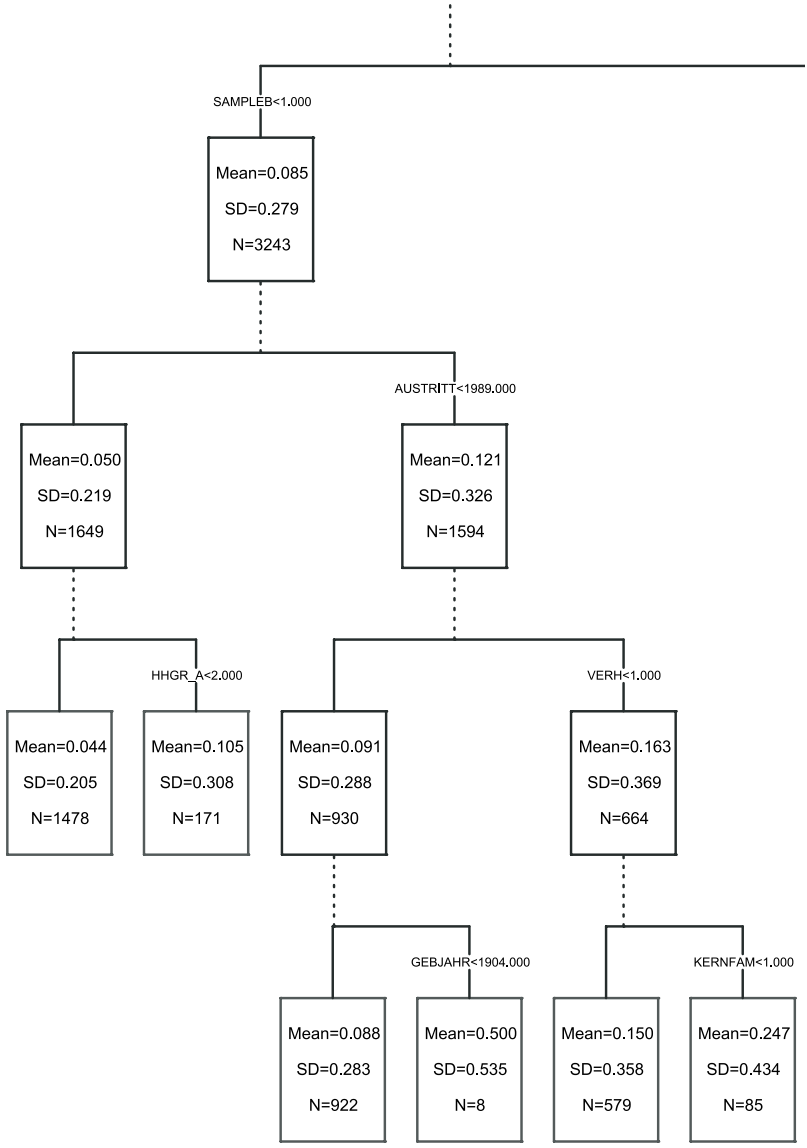
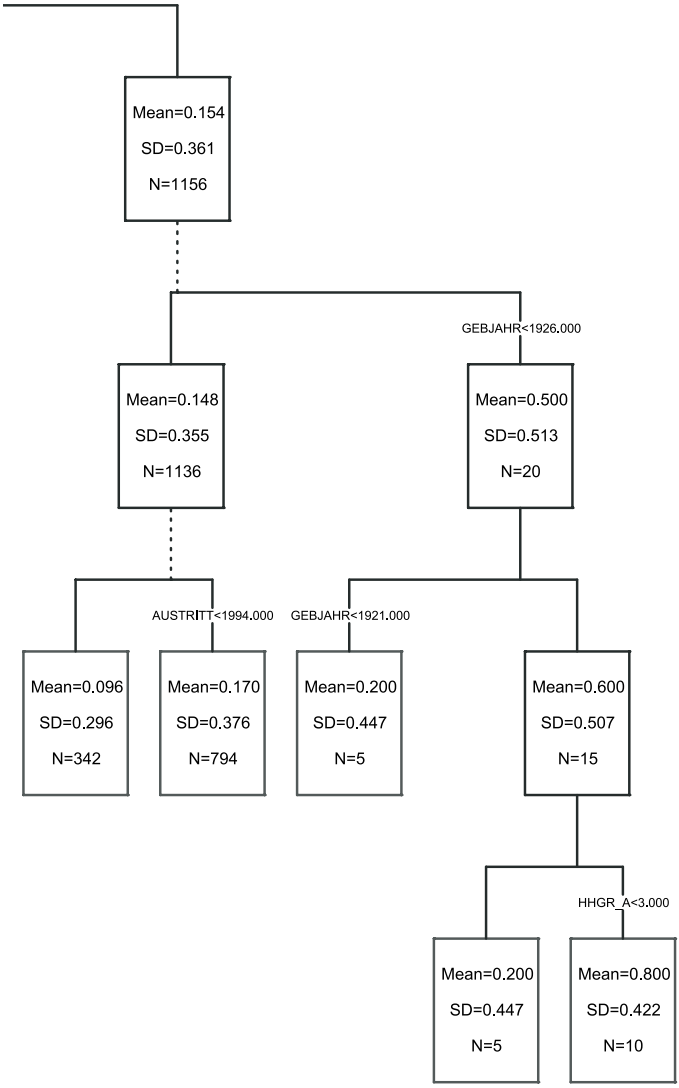


Abb. 4: Klassifikationsbaum zur Erklärung der Nichtauffindbarkeit im Melderegister


Mean=0.154
SD=0.361
N=1156
GEBJAHR<1926.000
Mean=0.148
SD=0.355
N=1136
Mean=0.500
SD=0.513
N=20
AUSTRITT<1994.000
GEBJAHR<1921.000
Mean=0.096
SD=0.296
N=342
Mean=0.170
SD=0.376
N=794
Mean=0.200
SD=0.447
N=5
Mean=0.600
SD=0.507
N=15
HHGR_A<3.000
Mean=0.200
SD=0.447
N=5
Mean=0.800
SD=0.422
N=10

(Deutsche (West)), die vor 1989 austraten, sind vor allen Dingen Personen, die unverheiratet sind (16,3 Prozent von n=664), schlecht auffindbar, vor allem wenn sie nicht zur Kernfamilie des Haushaltsvorstands gehören (24,7 Prozent von n=85). Bei den nach 1989 Ausgetretenen werden vor allem Einpersonenhaushalte schlecht wieder aufgefunden (10,5 Prozent, n=171). In Abbildung 4 ist der komplette Baum dargestellt.

Etwas vereinfacht lassen sich also vor allem Ausländer und Unverheiratete schlechter auffinden. Für beide Subgruppen sind systematische Zusammenhänge zu erhöhter Mortalität denkbar.

So haben in Deutschland lebende Ausländer zwar nicht generell eine geringere Lebenserwartung als Deutsche[23] , bei den nicht mehr auffindbaren Ausländern könnte es sich aber vorzugsweise um Rückkehrer in die Heimatländer handeln. Falls diese Rückkehr in Zusammenhang mit Krankheit oder Tod stehen sollte, wäre dies eine mögliche Quelle deutlicher Verzerrung bei der Schätzung ihrer Lebenserwartung. Zu dieser Vermutung passt, dass die verbliebenen Ausländer im SOEP und in der Verbleibstudie eine ausgesprochen hohe Lebenserwartung aufweisen[24]. Dies erklärt allerdings nicht die Unterschiede zwischen SOEP und amtlicher Statistik, da auch in der amtlichen Statistik Lebensjahre von Migranten nach einer Remigration nicht mehr erfasst werden. Somit ist auch die amtliche Statistik nicht zur Schätzung der Lebenserwartung von Migranten geeignet, solange der Verdacht nicht widerlegt ist, dass ein Zusammenhang zwischen Remigration ins Heimatland und der Sterbewahrscheinlichkeit besteht. Für die zweite in der Verbleibstudie schwer auffindbare Gruppe, unverheiratete Personen, ist allgemein gut belegt, dass sie deutlich erhöhte Sterberisiken gegenüber Verheirateten besitzen (Klein 1993).

Es erscheint daher nicht ausgeschlossen, dass die 775 nicht gefundenen Personen sowie die nicht in der Verbleibstudie berücksichtigten Personen ein Grund für die Unterschätzung des Sterberisikos in der SOEP-Stichprobe auch nach der Verbleibstudie sind.

11 Zusammenfassung

Die Infratest-Verbleibstudie hat keine schwerwiegenden Auswirkungen auf die Schätzung der Lebenserwartung mithilfe des SOEP. Die Werte der geschätzten Survivor-Funktionen vor und nach der Verbleibstudie unterscheiden sich bei bei-

23 Im Gegenteil ist ihre Lebenserwartung aus unterschiedlichen Gründen sogar höher, vgl. z.B. Klein (2005:99).

24 Auch weist Klein (2005:99) darauf hin, „[...] dass Ausländer bei ernsthafter Erkrankung nicht selten in ihr Heimatland zurückkehren."

den Geschlechtern zwar um bis zu 2,7 Prozentpunkte[25], die Verringerung der Lebenserwartung bei Berücksichtigung neuer Informationen über den Vitalstatus ist aber nicht signifikant.

Diese fehlende Verbesserung der Schätzung kann mehrere Ursachen besitzen. Mithilfe einer Simulation konnte gezeigt werden, dass bei einem Datensatz von der Größenordnung des SOEP relativ starke informative Zensierungsmechanismen notwendig sind, um signifikante Abweichungen zwischen Mortalitätsdaten mit und ohne informativen Zensierungsprozess zu erzielen. Der Grund hierfür liegt vor allem in der geringen Ereignisrate.

Ein weiterer Grund für die fehlende Verbesserung der Schätzung durch die Verbleibstudie könnte darin begründet liegen, dass diejenigen, die in die Verbleibstudie gelangten, eine relativ gesunde Subgruppe darstellten. In diesem Fall würde auch die Verbleibstudie zu einer Unterschätzung der Sterbewahrscheinlichkeit führen, aber kaum Unterschiede zum SOEP vor der Verbleibstudie aufweisen. Für eine Teilmenge der Ausgeschiedenen (diejenigen, für die die Verbleibstudie keine neuen Informationen über den Vitalstatus erbrachte) konnten zwei Subgruppen identifiziert werden, für die ein solcher Mechanismus des Ausscheidens morbider Personen aus der Verbleibstudie selbst plausibel ist: Ausländer und Unverheiratete.

Vergleicht man nicht die Unterschiede zwischen dem SOEP mit und dem SOEP ohne die Verbleibstudie, sondern die Sterbewahrscheinlichkeiten auf der Basis des SOEP mit denen der amtlichen Statistik, dann zeigen sich deutliche Unterschiede zwischen SOEP und amtlicher Statistik. Diese Unterschiede sind unabhängig davon, ob man die Gewichtung verwendet, für Frauen und Männer in der erwarteten Richtung signifikant. In jedem Fall wird mit dem SOEP die Sterbewahrscheinlichkeit unterschätzt.

Studien zur sozialen Differenzierung der Lebenserwartung wären in ihren Schlussfolgerungen von diesen Unterschieden dann betroffen, wenn die erhöhte Lebenserwartung der SOEP-Teilnehmer nicht für alle gesellschaftlichen Subgruppen gleichermaßen, sondern für manche stärker als für andere gälte. Genau dies wird aber durch die positive Korrelation zwischen Bildung und Teilnahmewahrscheinlichkeit einerseits, Bildung und Lebenserwartung andererseits nahegelegt. Daher vermuten wir eine Unterschätzung der differentiellen Lebenserwartung durch Schätzungen auf der Basis von Paneldaten. Die Klärung dieses Effekts muss weiteren Studien vorbehalten bleiben.

25 Beispielsweise wird die Wahrscheinlichkeit von Frauen, das 90. Lebensjahr zu erreichen, vor der Verbleibstudie auf 24,0 Prozent, nachher auf 21,3 Prozent geschätzt, bei Männern wird die Wahrscheinlichkeit, das 86. Lebensjahr zu erreichen vor der Verbleibstudie auf 24,6 Prozent, nachher auf 22,0 Prozent geschätzt.

Anhang I: Details der Simulation

Für die Fallzahl gilt: $n_{gesamt} = n_S + n_H = 2n_S = 2n_H$, für die Sterberisiken $h_S(t) = f \times h_H(t)$. h(t) ist die so genannte Hazardfunktion, die hier die Wahrscheinlichkeit angibt, innerhalb eines Jahres zu sterben, wenn man das Alter t erreicht hat. Die Hazardfunktionen $h_S(t)$ und $h_H(t)$ werden dabei so berechnet, dass im ersten Jahr $h_{gesamt}(t) = (n_S \times h_S(t) + n_H \times h_H(t)) / n_{gesamt}$ gilt. In der ersten Welle ist in jedem Alter das Sterberisiko in Population und Stichprobe gleich (kein Bias, kein sampling error). Für die Population unterstellen wir in jedem Jahr ein Sterberisiko, das dem der offiziellen Sterbetafeln für Männer 2001/03 entspricht. Der zusätzliche Gewinn einer realistischeren Modellierung wäre bei großem Aufwand äußerst gering, da das Ausmaß des Bias nicht stark von dieser Basisrate abhängt und da zudem Veränderungen der Sterberaten im Zeitverlauf sehr langsam vonstatten gehen. Um die Konsequenzen für (insbesondere die Signifikanz von) Schätzungen mithilfe des SOEP abschätzen zu können, ist es dagegen wichtig, die Stichprobengröße und die Altersverteilung in der Stichprobe entsprechend der Werte des SOEP zu modellieren. Die Simulation beginnt mit einer Stichprobe, die bezüglich der Fallzahl pro Geburtsjahrgang der Verteilung im SOEP entspricht. Um die Fallzahl relativ stabil zu halten, wie dies im SOEP bis 2000 der Fall war, werden vereinfachend jährliche Auffrischungen aus der Population im Umfang der Ausfälle modelliert (wobei bei diesen Auffrischungen zu berücksichtigen ist, dass auch die Population mit zunehmender Zeitdauer bei dieser Form der Modellierung zunehmend aus Gesunden besteht, da die Kranken häufiger sterben). Da die mittlere Bleibewahrscheinlichkeit in den bisherigen Wellen des SOEP bei etwa 0,94 lag, modellieren wir zunächst $p_S = 0{,}92$ und $p_H = 0{,}96$. Damit legen wir ein verdoppeltes Ausfallrisiko für „Kranke“ bei insgesamt etwa gleicher Ausfallrate wie im SOEP zugrunde. Es werden nun die erwarteten Sterbefälle über alle 18 Wellen kumuliert. Die Survivor-Funktion in der Stichprobe mit dem oben modellierten Ausfallprozess wird mit der Survivor-Funktion einer anfangs identisch zusammengesetzten Stichprobe aus derselben Population verglichen, in der es keine anderen als todesbedingte Ausfälle gibt. Anschließend wird ein Log-Rank-Test auf Gleichheit der beiden Survivor-Funktionen durchgeführt.

Anhang II: Details der Schätzung der Lebenserwartung

Für die Berechnung des Modells vor der Verbleibstudie 2001 werden alle „*dropouts*“ als im Austrittsjahr zensiert betrachtet. Es gab zwar bereits vor der Verbleibstudie für einige Dropouts eine Information über ihr Ableben nach dem Ausscheiden, doch ist diese aus der Feldarbeit stammende Information – wie die Ergebnisse der Verbleibstudie zeigen – unvollständig. Es konnte zwar für einige

Ausgeschiedene im Rahmen der Feldarbeit das definitive Ableben festgestellt werden, umgekehrt jedoch heißt das nicht, dass alle anderen Dropouts noch leben. Diese Annahme wäre naiv und würde das Ergebnis noch stärker verzerren. Die vorsichtigere Annahme (nach dem Ausscheiden liegt keine Kenntnis des Vitalstatus mehr vor: Die Fälle sind also ab dem Austrittsjahr zensiert) ist daher vorzuziehen. Für die Berechnung des Modells nach der Verbleibstudie müssen daher zwei Subgruppen unterschieden werden. Für diejenigen, die nicht in die Verbleibstudie einbezogen wurden, wird nach wie vor angenommen, dass sie im Austrittsjahr zensiert sind.[26] Für die 7 127 Fälle, für die in der Verbleibstudie der Vitalstatus geklärt werden konnte, kann nun je nach Ergebnis der Verbleibstudie das recherchierte Todesjahr als Zeitpunkt des Ereigniseintritts (Tod) genutzt werden oder ein Überleben bis 2001 und damit eine Zensierung erst im Jahr 2001 angenommen werden. Beide Modelle wurden verglichen. Dabei bestehen die beiden Modelle zu einem großen Teil (83 Prozent) aus exakt identischen Episoden. Betrachten wir jedes Personenjahr als Episode, so gibt es vor der Verbleibstudie 307 660, nach der Verbleibstudie 370 548 Episoden, in denen die 307 660 komplett enthalten sind.[27]

Literatur

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984): Classification and Regression Trees. Wadsworth, Belmont.

Dufuoil, C., Brayne, Carol und Clayton, D. (2004): Analysis of longitudinal studies with death and dropout: A case study. Statistics in Medicine, 23, 2215-26.

Finkelstein, D. M., Muzikansky, A. und Schoenfeld, D. A. (2003): Comparing survival of a sample to that of a standard population. In: Journal of the National Cancer Institute, 95, 1434-1439.

Fuchs, J. und Hansmeier, T. (1996): Ein Krankheitsersatzindex: Konstruktion und Validierung. Sozial- und Präventivmedizin, 41, 231-239.

Heller, G. und Schnell, R. (2000): The Choir Invisible. Zur Analyse der gesundheitsbezogenen Panelmortalität im SOEP; in: Helmert, U.; Bamman, K.; Voges, W.; Müller, R. (Hrsg.): Müssen Arme früher sterben? Soziale Ungleichheit und Gesundheit in Deutschland, München (Juventa), S.115-134.

26 Für mehr als die Hälfte der Ausgeschiedenen gilt damit nach wie vor, dass wir die Annahme uninformativer Zensierung treffen müssen, damit die Schätzer unverzerrt sind. Besonders problematisch wäre es, wenn die im Melderegister nicht Auffindbaren systematisch andere Sterberisiken hätten als die Übrigen.

27 Es würde nahe liegen, diese identischen Episoden aus der zweiten Analyse auszuschließen und hier nur die 62 888 neuen Episoden mit den 307 660 alten Episoden zu vergleichen, da dann deutlichere Differenzen zu erwarten wären, doch erlaubt dieses Vorgehen nicht, den Fehler zu identifizieren, den man ohne Berücksichtigung der Verbleibstudie begeht.

Infratest Sozialforschung (2002): Verbesserung der Datengrundlage für Mortalitäts- und Mobilitätsanalysen: Verbleibstudie bei Panelausfällen im SOEP. München. http://www.diw.de/deutsch/sop/service/doku/docs/verbleibstudie.pdf

Kaplan, E.L. und Meier, P. (1958): Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.

Klein, Th. (1993): Familienstand und Lebenserwartung. In: Zeitschrift für Familienforschung, 5, 99-114.

Klein, Th. (2005): Sozialstrukturanalyse: Eine Einführung. Reinbek bei Hamburg: Rowohlt.

Klein, Th. und Unger, R. (2001): Einkommen, Gesundheit und Mortalität in Deutschland, Großbritannien und den USA. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 53, 96-110.

Panneberg, M. u.a. (2003): Sampling and Weighting. In: John P. Haisken-DeNew und Joachim R. Frick: Desktop Companion to the German Socio-Economic Panel Study (SOEP). Version 7, 137-169. DIW Berlin.

Rendtel, U. (1995): Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität. Frankfurt: Campus.

Singer, J. D. und Willett, J. B. (2003): Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence. Oxford University Press.

SOEP Group (2001): The German Socio-Economic Panel (GSOEP) after more than 15 years – Overview. In: Holst, E., Lillard, D. R. und DiPrete, Th. A. (Hrsg.): Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP 2000), Vierteljahreshefte zur Wirtschaftsforschung, 70, 1, 7-14.

Spieß, M. und Kroh, M. (2004): Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panels (GSOEP)(1984 until 2003). Berlin: DIW Research Notes 28a.

Statistisches Bundesamt (Hrsg.)(2004): Periodensterbetafeln für Deutschland: Allgemeine und abgekürzte Sterbetafeln von 1871/1881 bis 2001/2003. Wiesbaden: Statistisches Bundesamt.

Unger, R. (2003): Soziale Differenzierung der aktiven Lebenserwartung im internationalen Vergleich. Wiesbaden: Deutscher Universitäts-Verlag.

Ausfallgründe bei zufallsgenerierten Telefonstichproben am Beispiel des Gabler-Häder-Designs

Nina Baur

Zusammenfassung

In zwei 2005 durchgeführten CATI-Studien wurden mit dem Gabler-Häder-Design generierte Stichproben verwendet und verschiedene Ausfallgründe detailliert erfasst. Diese Ausfallgründe werden systematisch und mit denen früherer Studien verglichen. Obgleich die Fallzahl der untersuchten Studien sehr gering ist, lassen sich folgende Ergebnisse vorläufig festhalten: Der konstant wichtigste Verweigerungsgrund ist Desinteresse an der betreffenden Studie. Seit 2001 fallen weiterhin immer mehr Befragte aufgrund von Zeitknappheit oder Misstrauen gegenüber der Studie aus. Da die meisten Zielpersonen das Interview gleich zu Beginn der Kontaktaufnahme abbrechen, werden ein gut formulierter Gesprächseinstieg, qualifizierte Interviewer und – nach Möglichkeit – Warmkontakte immer wichtiger. Des Weiteren sind – qualitative – Studien über die Einstellungs- und Verhaltensunterschiede zwischen Verweigerern und Studienteilnehmern dringend erforderlich, um das Ausmaß der durch Verweigerungen entstehenden systematischen Fehler besser abschätzen zu können.

1 Einleitung

Das Gabler-Häder-Design gilt derzeit als das beste Verfahren, um für Deutschland zufallsgenerierte Telefonstichproben zu ziehen (vgl. hierzu Gabler / Häder 1999; Häder / Gabler 2000; Deutschmann / Häder 2002). Gabler und Häder (in diesem Band) diskutieren Probleme, die dadurch entstehen, dass Zielpersonen unter Umständen gar nicht erst in die Bruttostichprobe gelangen, weil sie keinen Telefonanschluss besitzen, weil sie einen anonymen Anschluss haben oder weil sie nur noch mobil oder über einen Internettelefonanschluss erreichbar sind. Vorgestellt werden auch Pläne für einen neuen Designtyp, um diese Probleme zu beheben. Undercoverage kann auch durch Totalausfälle (= unit-nonresponse) entstehen, also dadurch, dass ein Haushalt bzw. eine Person zwar in die Bruttostichprobe, nicht aber in die Nettostichprobe gelangt (Behnke et al. 2006).

Ausgehend von der Annahme, dass mit Hilfe des Gabler-Häder-Designs eine bestmögliche Bruttoausgangsstichprobe von Telefonnummern gezogen wurde,

widmet sich der folgende Beitrag den Ausfallgründen bei zufallsgenerierten Telefonstichproben. Hierzu werden zunächst die verwendeten Daten beschrieben. Anschließend wird die Entwicklung der stichprobenneutralen Ausfälle und der Totalausfälle erläutert. Letztere lassen sich unterteilen in Nichterreichbarkeit und Verweigerungen. Aufbauend auf diesen Ergebnissen schließt der Beitrag mit Empfehlungen für die Forschungspraxis.

2 Beschreibung der Studien

Der Beitrag vergleicht die Ergebnisse zweier Studien aus dem Jahr 2005, bei denen die Ausfallgründe detailliert erfasst wurden, mit den Ausfallgründen früherer Studien. Bei den Studien aus dem Jahr 2005 handelt es sich um zwei an der Katholischen Universität Eichstätt-Ingolstadt durchgeführten CATI-Erhebungen:

- *Die Studie „Hartz IV im Spiegel der Bevölkerungsmeinung" (im Folgenden Studie „Hartz IV")* befasste sich mit Einstellungen zur Arbeitslosigkeit und zum Sozialstaat, insbesondere der Arbeitslosenversicherung. Sie wurde von Siegfried Lamnek und der Universität Eichstätt-Ingolstadt finanziert. Die Projektleitung hatten Siegfried Lamnek und Nina Baur inne. Die Datenerhebung fand zwischen dem 7. und 22. März 2005 statt.
- *Die Studie „Die Qualität von tagesaktuellen Printmedien aus der Publikumsperspektive. Theoretische Überlegungen und empirische Untersuchung" (im Folgenden Studie „Tageszeitungen")* wurde von der DFG finanziert. Projektleiter war Klaus Arnold. Die Datenerhebung fand zwischen dem 26. April und dem 23. Juni 2005 statt.

In beiden Fällen zog Siegfried Gabler (ZUMA) die Stichprobe nach dem Gabler-Häder-Design, wofür wir ihm an dieser Stelle herzlich danken möchten. Die Zielperson wurde mit Hilfe der Last Birthday-Methode ermittelt. Die Feldphase einschließlich der Interviewerschulungen und der Supervision des CATI-Labors organisierte Nina Baur. Bei den Interviewern handelte es sich um gut geschulte Studierende der Universität Eichstätt-Ingolstadt. Die Studien sind also einerseits gut vergleichbar, weil die Erhebungsmethoden sich ähneln, andererseits variieren sie u. a. in der Erhebungsdauer (ca. 2 Wochen vs. ca. 8 Wochen) und der Fragebogenlänge.

Die Daten aus früheren Studien wurden anderen Publikationen zur Stichprobenproblematik bei zufallsgenerierten Telefonstichproben mit dem Gabler-Häder-Design entnommen. Vergleiche der Studien aus dem Jahr 2005 mit diesen Studien sind mit Vorsicht zu interpretieren:

1) Die Zahl der Studien ist insgesamt sehr gering, so dass unklar ist, ob Unterschiede zwischen den Ausfallquoten und -gründen in den Studien von 2005 und in früheren Studien auf zufällige Schwankungen, Institutseffekte oder reale Veränderungen im Befragtenverhalten zurückzuführen sind.

2) Jeder Forscher ordnet Ausfälle anders zu. So klassifiziert Häder (2000: 10) auch Freizeichen, Belegtzeichen bei 10 Kontaktversuchen, Fax- und Modem-Karten sowie Anrufbeantworter als stichprobenneutral. Ich ordne diese Ausfälle als potenziell verzerrend ein: Freizeichen, Belegtzeichen und Anrufbeantworter weisen auf eine extrem schwierig zu erreichende Klientel hin. Je nach Forschungsfrage kann der Ausfall dieser Personengruppe das Ergebnis beeinflussen. Faxgeräte können auch gleichzeitig Telefonanschlüsse sein.
3) Ich habe die Vergleichzahlen aus Veröffentlichungen anderer Autoren zur Stichprobenproblematik entnommen. Teilweise sind nur Prozentangaben angegeben, so dass ich die Fallzahlen rückrechnen musste.

Zum Vergleich werden zusätzlich die Ausfallstatistiken zweier Studien dargestellt, deren Stichproben mit Hilfe des Einwohnermeldeamtsregisters gezogen wurden. Blasius und Reuband (1995) haben in ihrer Studie Anfang der 1990er demonstriert, dass bei sorgfältiger Stichprobenpflege eine Ausschöpfungsquote von etwa 95% erreicht werden kann. Dies ist als vorbildlich und gleichzeitig als Maximum anzusehen, was überhaupt erreicht werden kann. Der ALLBUS eignet sich deshalb als Beispiel, weil auch sein Stichprobendesign als vorbildlich gilt und darüber hinaus gut dokumentiert ist. Beim ALLBUS wurde persönlich-mündlich befragt (Schneekloth / Leven 2003).

3 Stichprobenneutrale Ausfälle

Bei RLD-Verfahren – auch beim Gabler-Häder-Design – werden zum Teil Nummern generiert, die gar nicht existieren (Gabler / Häder 1997; Gabler / Häder 1998, Häder / Gabler 1998), so dass stichprobenneutrale Ausfälle entstehen. Tabelle 1 zeichnet die Entwicklung stichprobenneutraler Ausfälle zwischen 1999 und 2005 bei den untersuchten Studien nach.

Am Beispiel dieser Studien wird deutlich, dass Ende der 1990er rund ein Drittel der mit Hilfe des Gabler-Häder-Designs generierten Telefonnummern als stichprobenneutrale Ausfälle zu klassifizieren waren, weil die Nummern nicht existierten oder weil die erreichten Personen nicht zur Zielpopulation gehörten (Geschäftsanschlüsse, falsche Altersgruppe etc.). Ende der 1990er erhöhte die Telekom die Zahl der Nummernblöcke, ohne dass die Zahl der Telefonnummern entsprechend stieg. Die einzelnen Blöcke sind seitdem schlechter besetzt, so dass sich die Zahl der nichtexistierenden Nummern bei einer Gabler-Häder-Stichprobe und damit die „vergeblichen“ Wahlversuche erhöht haben (Gabler / Schürle 2002). Entsprechend liegen die stichprobenneutralen Ausfälle heute bei rund 45%, was die Erhebungskosten deutlich erhöht.

Tabelle 1: Stichprobenneutrale Ausfälle in verschiedenen Studien

Erhebungszeitraum	1991 / 1992		3.2. – 13.3.1999		April 1999		14.2. – 23.3.2000		9.10. – 1.11.2000	
Region	Köln		Mannheim		BRD		Mannheim		BRD	
Studie	Sozialer und kultureller Wandel		Lebensstile in Mannheim		ZUMA-Stichproben-experiment		Lebensführung in sozialen Netzwerken		Medien-nutzung	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Telefonisch		Telefonisch	
Stichprobenverfahren	Einwohner-meldeamt		Gabler-Häder		Gabler-Häder		Gabler-Häder		Gabler-Häder	
Bruttoausgangsstichprobe	**469**	**100,0%**	**5.000**	**100,0%**	**1.500**	**100,0%**	**4.397**	**100,0%**	**36.960**	**100,0%**
Ansage „kein Anschluss unter dieser Nummer“			1.171	23,4%	399	26,6%	1.042	23,7%	13.277	35,9%
Ansage „Rufnummer geändert“ und Ähnliches			26	0,5%	4	0,3%	94	2,1%	105	0,3%
Geschäftsadresse			288	5,8%	78	5,2%	233	5,3%	2.119	5,7%
Falsches Bundesland / falscher Ort			24	0,5%			33	0,8%		
Stichprobenneutrale Ausfälle	**28**	**6,0%**	**1.509**	**30,2%**	**481**	**32,1%**	**1.402**	**31,9%**	**15.501**	**41,9%**
Bereinigte Bruttostichprobe	**441**	**94,0%**	**3.491**	**69,8%**	**1.019**	**67,9%**	**2.995**	**68,1%**	**21.459**	**58,1%**
Quelle	Blasius / Reuband 1995: 71		Otte 2002: 88-90		Häder 2000: 10		Otte 2002: 88-90		Deutschmann / Häder 2002: 67	

Telefonstichproben nach dem Gabler-Häder-Design „können für bundesweite Umfragen, Umfragen in einzelnen Vorwahlbereichen und ausgewählten Gemeinden bereitgestellt werden. Bei letzteren ist allerdings zu berücksichtigen, dass Vorwahlbereichs- und Gemeindegrenzen nicht kongruent sind und die Telefonnummern innerhalb eines Vorwahlbereiches auch nicht nach Gemeinden vergeben werden.“ (Häder 2000: 8). Wie die Tabelle zeigt, ist die Zahl der Fehlklassifikationen sehr gering: Nummern von Orten bzw. Bundesländern, die nicht zur Zielregion gehören, wurden in deutlich weniger als 1% der Fälle generiert.

Tabelle 1: Stichprobenneutrale Ausfälle in verschiedenen Studien *(Fortsetzung)*

Erhebungszeitraum	**5.2. – 14.3.2001**		**März 2005**		**April – Juni 2005**		**2000**		**2000**	
Region	Mannheim		Bremen, Baden-Württemberg, NRW, Sachsen-Anhalt		BRD		Westdeutschland		Ostdeutschland	
Studie	Image der Stadt Mannheim		Hartz IV		Qualität von Tageszeitungen		ALLBUS		ALLBUS	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Face-to-Face		Face-to-Face	
Stichprobenverfahren	Gabler-Häder		Gabler-Häder		Gabler-Häder		Einwohnermeldeamt		Einwohnermeldeamt	
Bruttoausgangsstichprobe	**5.000**	**100,0%**	**13.040**	**100,0%**	**20.000**	**100,0%**	**~ 5.106**	**100,0%**	**~ 2.346**	**100,0%**
Leitung tot			783	6,0%	1.332	6,7%				
Ansage „kein Anschluss unter dieser Nummer“	1.486	29,7%	4.601	35,3%	7.307	36,5%				
Ansage „Rufnummer geändert“ und Ähnliches										
Geschäftsadresse	289	5,8%	470	3,6%	735	3,7%				
Falsches Bundesland / falscher Ort	26	0,5%	41	0,3%						
Nur Personen in falscher Altersgruppe im Haushalt			11	0,1%						
Sonstige Out of Frame			5	< 0,1%	1	< 0,1%				
Stichprobenneutrale Ausfälle	**1.801**	**36,0%**	**5.911**	**45,3%**	**9.375**	**46,9%**	**ca. 756**	**14,8%**	**ca. 293**	**12,5%**
Bereinigte Bruttostichprobe	**3.199**	**64,0%**	**7.129**	**54,7%**	**10.625**	**53,1%**	**~ 4.350**	**85,2%**	**~ 2.052**	**87,5%**
Quelle	Otte 2002: 88-90		Lamnek / Baur		Arnold		Schneekloth / Leven 2003: 23-34		Schneekloth / Leven 2003: 23-34	

4 Nichtkontaktraten

Im Gegensatz zu stichprobenneutralen Ausfällen stellt Undercoverage infolge von Totalausfällen für die Stichprobenqualität insofern ein Problem dar, dass hierdurch Ziel- und Auswahlgesamtheit nicht mehr übereinstimmen. Die induktive Statistik taugt in einem solchen Fall nur noch eingeschränkt als Verallgemeinerungsstrategie, da sie sich auf die Auswahl- und nicht auf die Zielgesamtheit bezieht (Behnke et al. 2006).

Seit Anfang der 1970er steigt die Nonresponserate bei Umfragen kontinuierlich an (Schnell 1997: 91). Während in aus dem Telefonbuch gezogenen Stichproben die Ausschöpfungsquoten Anfang der 1990er zwischen 26% und 60% schwankten, betrugen diese bei Studien mit RDD- bzw. RLDD-Stichproben 60% bis 70% (Blasius / Reuband 1995: 67). Ich vermute allerdings, dass diese Ausschöpfungsquoten etwas zu hoch gegriffen sind, da in der Regel Freizeichen, Belegtzeichen, Faxgeräte und Anrufbeantworter als „neutrale" Ausfälle klassifiziert und deshalb nicht als Ausfälle gezählt werden. Solche Ausfallgründe können je nach Fragestellung die Stichprobe aber durchaus verzerren (wenn z. B. gerade die Unterschiede zwischen extrem mobilen und weniger mobilen Personen interessieren).

Die sinkenden Ausschöpfungsquoten sind insofern problematisch, da empirisch bestätigt ist, dass sie i. d. R. mit systematischen Verzerrungen einhergehen. Allerdings können auch „schlecht" ausgeschöpfte Stichproben gute Qualität aufweisen (Schneekloth / Leven 2003: 49-50).

Wichtiger als die Ausschöpfungsquote an sich sind also die Ausfallgründe und ob diese mit dem Untersuchungsziel zusammenhängen. Aus diesem Grund unterscheidet Häder (1994) zwischen der Nichtkontaktrate und sonstigen Ausfällen, die im weitesten Sinne als Verweigerungen zu zählen sind. „Die Nichtkontaktrate gibt Auskunft darüber, zu welchem Anteil potentielle Untersuchungsteilnehmer nicht lokalisiert bzw. tatsächlich erreichbar sind. Bei Telefonumfragen führen besonders maschinelle Anrufbeantworter, Besetztzeichen und unterbrochene Anschlüsse zur Vergrößerung der Zahl nichthergestellter Kontakte." (Häder 1994: 19).

Haushalte sind (unabhängig von der Erhebungsform) immer schlechter zu erreichen. Die Kontaktrate stabilisierte sich dann Ende der 1980er Jahre u. a. durch eine verstärkte Bemühung der Institute und eine Lockerung der Zufallsauswahl auf der vorletzten Stufe des ADM-Designs (Schnell 1997: 92-106). Die Kontaktrate schwankt außerdem jahreszeitlich. So ist sie etwa in den Oster- und Sommerferien niedriger (Stögbauer 2000: 95). Mitte der 1990er wurden bei telefonischen Befragungen meist etwa doppelt so viele Haushalte bzw. Zielpersonen *nicht* erreicht wie bei persönlichen Befragungen (Schnell 1997: 116-119). Dies bestätigt sich, wenn man neuere Studien betrachtet. Wie Tabelle 2 zeigt, liegt die Nichtkontaktrate bei den zum Vergleich herangezogenen persönlich-mündlichen Befragungen bei unter 10%. Bei Telefonumfragen mit dem Gabler-Häder-Design

wurden dagegen um die Jahrtausendwende zwischen 20% und 30% der Zielhaushalte nicht erreicht. In den Eichstätter Studien aus dem Jahr 2005 wurden dagegen 54% bis 60% der Zielhaushalte nicht erreicht.

Tabelle 2: Nichtkontaktrate in verschiedenen Studien

Erhebungszeitraum	**1991 / 1992**		**3.2. – 13.3.1999**		**April 1999**		**14.2. – 23.3.2000**		**9.10. – 1.11.2000**	
Region	Köln		Mannheim		BRD		Mannheim		BRD	
Studie	Sozialer und kultureller Wandel		Lebensstile in Mannheim		ZUMA-Stichproben-experiment		Lebensführung in sozialen Netzwerken		Medien-nutzung	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Telefonisch		Telefonisch	
Stichprobenverfahren	Einwohner-meldeamt		Gabler-Häder		Gabler-Häder		Gabler-Häder		Gabler-Häder	
Bereinigtes Brutto	**441**	**100,0%**	**3.491**	**100,0%**	**1.019**	**100,0%**	**2.995**	**100,0%**	**21.459**	**100,0%**
Haushalt nicht erreicht, weil ...										
Nummer vorübergehend nicht erreichbar			29	0,8%			17	0,6%		
Freizeichen			333	9,5%	218	21,4%	173	5,8%	1.850	8,6%
Besetzt			194	5,6%	33	3,2%	304	10,2%	149	0,7%
Anrufbeantworter			256	7,3%	73	7,2%	57	1,9%	528	2,5%
ISDN-Modem-Karte / Faxgerät			235	6,7%	6	0,6%	323	10,8%	1.973	9,2%
Unterbrechung der Verbindung			15	0,4%			2	0,1%		
Verzerrende Ausfälle I: Nichtkontaktrate	**40**	**9,9%**	**1.062**	**30,4%**	**330**	**32,4%**	**876**	**29,2%**	**4.500**	**21,0%**
Kontaktrate	**401**	**90,1%**	**2.429**	**69,6%**	**689**	**67,6%**	**2.119**	**70,8%**	**16.959**	**79,0%**
Quelle	Blasius / Reuband 1995: 71		Otte 2002: 88-90		Häder 2000: 10		Otte 2002: 88-90		Deutschmann / Häder 2002: 67	

Schnell (1997: 116-119) nennt als Hauptursache für die niedrigeren Kontaktraten bei Telefonumfragen die meist wesentlich kürzeren Feldzeiten bei Telefonumfragen (durchschnittlich vier Tage vs. 2 Monate). Dafür spricht, dass die Studie „Tageszeitung“ mit einer im Vergleich zur Studie „Hartz IV“ wesentlich *längeren Feldzeit* (etwas über acht Wochen vs. etwas über zwei Wochen) auch eine um etwa

5 Prozentpunkte niedrigere Nichtkontaktrate aufweist. Dies erklärt allerdings noch nicht die großen Unterschiede der Kontaktraten zwischen früheren Studien.

Tabelle 2: Nichtkontaktrate in verschiedenen Studien *(Fortsetzung)*

Erhebungszeitraum	**5.2. – 14.3.2001**		**März 2005**		**April – Juni 2005**		**2000**		**2000**	
Region	Mannheim		Bremen, Baden-Württemberg, NRW, Sachsen-Anhalt		BRD		Westdeutschland		Ostdeutschland	
Studie	Image der Stadt Mannheim		Hartz IV		Qualität von Tageszeitungen		ALLBUS		ALLBUS	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Face-to-Face		Face-to-Face	
Stichprobenverfahren	Gabler-Häder		Gabler-Häder		Gabler-Häder		Einwohnermeldeamt		Einwohnermeldeamt	
Bereinigtes Brutto	**3.199**	**100,0%**	**7.129**	**100,0%**	**10.625**	**100,0%**	**~4.350**	**100,0%**	**~2.052**	**100,0%**
Haushalt nicht erreicht, weil ...										
Nummer vorübergehend nicht erreichbar	17	0,5%	47	0,7%	40	0,4%				
Freizeichen	218	6,8%	2.492	35,0%	3.310	31,2%				
Besetzt	435	13,6%	196	2,7%	168	1,6%				
Anrufbeantworter	94	2,9%	1.075	15,1%	1.431	13,5%				
ISDN-Modem-Karte / Faxgerät	217	6,8%	449	6,3%	826	7,8%				
Unterbrechung der Verbindung	5	0,2%	1	< 0,1%	4	< 0,1%				
Verzerrende Ausfälle I: Nichtkontaktrate	**986**	**30,8%**	**4.260**	**59,8%**	**5.779**	**54,4%**	**~231**	**5,3%**	**~29**	**1,4%**
Kontaktrate	**2.213**	**69,2%**	**2.869**	**40,2%**	**4.846**	**45,6%**	**~4.119**	**94,7%**	**~2.023**	**98,6%**
Quelle	Otte 2002: 88-90		Lamnek / Baur		Arnold		Schneekloth / Leven 2003: 23-34		Schneekloth / Leven 2003: 23-34	

Um die Kontaktwahrscheinlichkeit zu erhöhen, wird weiterhin empfohlen, die *Kontaktversuche über Wochentage und Tageszeiten zu variieren* (Porst et al. 1998: 14; 21). Dies wurde bei beiden Studien gemacht: Die Interviewzeiten waren Montag bis Freitag 14.00 bis 21.00 Uhr und am Wochenende 10.00 bis 21.00 Uhr. Auf Anfrage der Zielpersonen wurden einzelne Interviews auch auf den Vormittag verschoben. Tatsächlich war die Variation der Interviewzeiten bei der Studie „Tageszeitungen“ sogar geringer, weil es sich bei den Interviewern um Studierende han-

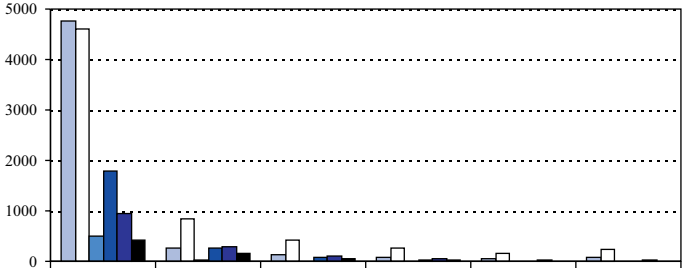
delte. Aufgrund der langen Feldphase und der Erhebungszeit während des Semesters konnte das CATI-Labor nachmittags und am Wochenende bisweilen nicht besetzt werden. Offensichtlich war die Variation dennoch groß genug, da die Kontaktrate bei dieser Studie dennoch höher war als bei der Studie „Hartz IV".

Weiterhin kann die Kontaktrate deutlich erhöht werden, wenn *mehrere Kontaktversuche* unternommen werden. Frühere Methodenexperimente zeigen, dass mindestens zwei Kontaktversuche unternommen werden sollten, die ideale Kontaktzahl aber fünf bis sechs beträgt, damit (für die meisten Fragestellungen) keine nennenswerten Verzerrungen mehr in der Zusammensetzung der Stichprobe bezüglich soziodemographischer Merkmale, Einstellungen und Verhaltensweisen auftreten. Mehr als sechs Kontaktversuche sind nur für extrem mobile Zielpersonen erforderlich (Porst et al. 1998: 15; Hüfken 2000). Hierbei handelt es sich hauptsächlich um junge Personen und besser Gebildete. Letztere verweigern aber selten, so dass diese leicht überrepräsentiert werden, wenn jede Nummer mehr als sechsmal angewählt wird (Blasius / Reuband 1995: 84-85). Für die meisten Fragestellungen genügen also fünf bis sechs Kontaktversuche. Hierin liegt möglicherweise auch ein Grund für die geringeren Kontaktraten bei den Eichstätter Studien: Wie Abbildung 1 und 2 auf S. 168 zu entnehmen ist, wurden die meisten Nummern maximal zweimal angewählt.

Ansonsten bestätigen sich die Befunde aus früheren Studien: Fast alle nicht-existierenden Leitungen werden beim ersten Wahlversuch identifiziert. Bei manchen nicht-existierenden Leitungen kommt keine Ansage der Telekom, sondern ein Belegtzeichen, das etwas anders klingt als das „normale" Belegtzeichen, das auf einen besetzten Anschluss hinweist. Erfahrene Interviewer und Supervisoren können diese Töne voneinander unterscheiden. Hierauf ist auch zurückzuführen, dass nicht alle toten Leitungen mit dem ersten Wahlversuch identifiziert wurden.

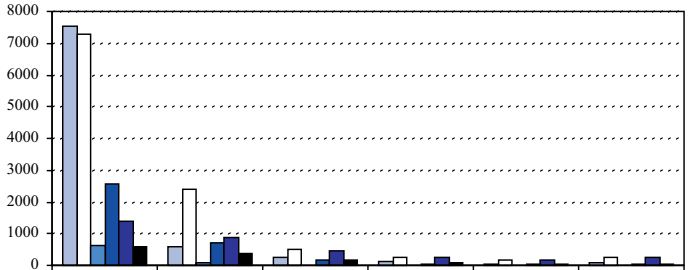
Etwa ein Drittel der Haushalte wird beim ersten Kontaktversuch nicht erreicht. Von den Nicht-Erreichten wird auch beim zweiten Kontaktversuch knapp unter die Hälfte nicht erreicht. Auch beim dritten Kontaktversuch reduziert sich die Zahl der Nicht-Erreichten noch einmal um die Hälfte. Danach stabilisiert sich bei der Studie „Tageszeitungen" die Zahl der Nicht-Erreichten, und es ist unklar, ob es sich hierbei um sehr schwer erreichbare Haushalte oder um nicht existierende, aber als solche nicht identifizierbare Anschlüsse handelt. Bei der Studie „Hartz IV" bleibt dagegen der Anteil der nicht erreichten Haushalte hoch, was wohl daran liegt, dass hier die Kontaktversuche innerhalb weniger Tage stattfanden (wegen der kürzeren Erhebungszeit) und dass die Erhebung in die Osterferien hineinreichte. Daraus lässt sich schließen, dass es – für die Erreichbarkeit – vorteilhaft ist, eine möglichst lange Erhebungszeit zu planen und erst einmal alle Nummern abzuwählen, um den Abstand zwischen den Kontaktversuchen möglichst groß zu halten (wobei zu beachten ist, dass lange Erhebungszeiten wieder andere metho-

dische Probleme mit sich bringen, z. B. mögliche Stimmungsumschwünge während der Erhebungszeit).



	1 (n=13.040)	2 (n=1.856)	3 (n=767)	4 (n=430)	5 (n=241)	>5 (n=366)
Leitung existiert nicht	4773	276	124	85	46	80
Haushalt nicht erreicht	4603	844	410	251	163	245
Out of Frame	490	32	2	2	0	1
Verweigerung	1801	253	66	23	8	6
Terminvereinbarung	944	297	115	43	15	29
vollständiges Interview	429	154	50	26	9	5

Abb. 1: Ergebnis von Kontaktversuchen (Studie „Hartz IV“)



	1 (n=20.000)	2 (n=5.052)	3 (n=1.548)	4 (n=775)	5 (n=422)	>5 (n=677)
Leitung existiert nicht	7553	577	266	117	44	83
Haushalt nicht erreicht	7277	2398	504	259	158	232
Out of Frame	617	100	13	4	1	1
Verweigerung	2585	719	153	62	30	41
Terminvereinbarung	1399	893	458	252	156	272
vollständiges Interview	569	365	154	81	33	48

Abb. 2: Ergebnis von Kontaktversuchen (Studie „Tageszeitungen“)

Es zeigt sich auch, dass sich die zusätzlichen Kontaktversuche durchaus lohnen: Bei den Haushalten, die erreicht werden, ist beim ersten Kontaktversuch der Anteil der Verweigerer wesentlich größer als bei späteren Kontaktversuchen. Anders

ausgedrückt: Schwer erreichbare Haushalte verweigern eher seltener als leicht erreichbare. Allerdings lassen die Daten diesbezüglich auch eine andere Interpretation zu: Es fällt auf, dass bei mehr Kontaktversuchen der Anteil der offenen Terminvereinbarungen immer größer wird, wobei es sich hauptsächlich um Personen handelt, die trotz Terminvereinbarung nicht zu Hause waren. Es könnte sich hierbei um eine Verweigerungsstrategie extrem mobiler Personen handeln – indem sie zum vereinbarten Termin einfach nicht da sind.

Auch in einem anderen Sinne ist es wünschenswert, mehrere Kontaktversuche zu unternehmen: Frühere Studien haben gezeigt, dass sich leicht und schwer erreichbare Personen im Lebensstil deutlich unterscheiden, so dass die Stichprobe erheblich verzerrt sein kann, wenn letztere unterrepräsentiert sind. Zu den schlecht Erreichbaren gehören Single-Haushalte, Vollzeit-Erwerbstätige, jüngere Menschen zwischen 20 und 30, Großstadtbewohner (aufgrund ihrer größeren Mobilität). Leichter zugänglich sind dagegen z. B. Hausfrauen und Arbeitslose (Schneekloth / Leven 2003: 29; Porst et al. 1998: 4; Schräpler 2000: 144; 147; Otte 2002: 94).

5 Verweigerungsrate

Von der Nichtkontaktrate zu unterscheiden sind Ausfälle, die im weitesten Sinne als Verweigerungen zu interpretieren sind. „Die Verweigerungsrate wird definiert als Verhältnis der Zahl derjenigen Personen, mit denen ein Kontakt zwar zustande gekommen ist, die aber das Interview nicht begonnen bzw. abgeschlossen haben, zur Gesamtzahl der Personen, mit denen ein Gesprächskontakt aufgenommen wurde." (Häder 1994: 19)

Wie in anderen Ländern erzielten Anfang und Mitte der 1990er auch in Deutschland Telefoninterviews eher niedrigere Response-Raten als Face-to-Face-Interviews. Im internationalen Vergleich verweigerten aber im Schnitt um etwa 10 Prozentpunkte weniger deutsche Haushalte die Teilnahme an Telefoninterviews (Schnell 1997: 119-126; Stögbauer 2000).

Wie Tabelle 3 zeigt, schwankte um die Jahrtausendwende die Verweigerungsrate bei CATI-Umfragen mit dem Gabler-Häder-Design zwischen 54% und 64%. Der niedrigere Wert entspricht in etwa der Verweigerungsrate westdeutscher Haushalte im ALLBUS 2000. Die Teilnahmebereitschaft ostdeutscher Haushalte lag beim ALLBUS 2000 noch um 10 Prozentpunkte höher. 2005 verweigerten dagegen etwa 75% der Haushalte die Teilnahme an der Studie. Um die Ursachen für die gestiegene Verweigerungsrate zu ergründen, ist es hilfreich, die Entwicklung der einzelnen Verweigerungsgründe zu analysieren. Auch für die Beurteilung der Stichprobenqualität ist die Differenzierung der Ausfallgründe bedeutend, da verschiedene Personengruppen aus unterschiedlichen Gründen die Teil-

nahme an Studien verweigern. In der Methodenliteratur werden i. d. R. folgende Ausfallgründe genannt:

Tabelle 3: Verweigerungsrate in verschiedenen Studien

Erhebungszeitraum	**1991 / 1992**		**3.2. – 13.3.1999**		**April 1999**		**14.2. – 23.3.2000**		**9.10. – 1.11.2000**	
Region	Köln		Mannheim		BRD		Mannheim		BRD	
Studie	Sozialer und kultureller Wandel		Lebensstile in Mannheim		ZUMA-Stichproben-experiment		Lebensführung in sozialen Netzwerken		Medien-nutzung	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Telefonisch		Telefonisch	
Stichprobenverfahren	Einwohner-meldeamt		Gabler-Häder		Gabler-Häder		Gabler-Häder		Gabler-Häder	
Zustande gekommene Kontakte	**401**	**100,0%**	**2.429**	**100,0%**	**689**	**100,0%**	**2.119**	**100,0%**	**16.959**	**100,0%**
Sprachprobleme	2	0,5%	114	4,7%	11	1,6%	98	4,6%	968	5,7%
Zielperson zu alt			14	0,6%	28	4,1%	13	0,6%		
Zielperson krank	7	1,7%	32	1,3%			18	0,8%		
Zielperson in Feldzeit nicht erreichbar			48	2,0%	24	3,5%	34	1,6%	1.101	6,5%
keine Zeit			104	4,3%	245	35,6%	51	2,4%		
kein Interesse			791	32,6%			894	42,2%		
Interview abgebrochen			26	1,1%	8	1,2%	16	0,8%		
Gatekeeper verweigert Zugang zur Zielperson			54	2,2%	3	0,4%	39	1,8%		
Gatekeeper legt wortlos auf			110	4,5%	34	4,9%	109	5,1%		
Anderer Ausfallgrund	30	7,5%	116	4,8%	50	7,3%	83	3,9%	8.779	51,8%
Verzerrende Ausfälle II: Verweigerungen u. a.	**39**	**9,7%**	**1.409**	**58,0%**	**403**	**58,5%**	**1.355**	**63,9%**	**19.848**	**64,0%**
Ausschöpfungsquote (realisierte Interviews)	**362**	**90,3%**	**1.020**	**42,0%**	**286**	**41,5%**	**764**	**36,1%**	**6.111**	**36,0%**
Quelle	Blasius / Reuband 1995: 71		Otte 002: 88-90		Häder 2000: 10		Otte 2002: 88-90		Deutschmann / Häder 2002: 67	

Tabelle 3: Verweigerungsrate in verschiedenen Studien *(Fortsetzung)*

Erhebungszeitraum	**5.2. – 14.3.2001**		**März 2005**		**April – Juni 2005**		**2000**		**2000**	
Region	Mannheim		Bremen, Baden-Württemberg, NRW, Sachsen-Anhalt		BRD		Westdeutschland		Ostdeutschland	
Studie	Image der Stadt Mannheim		Hartz IV		Qualität von Tageszeitungen		ALLBUS		ALLBUS	
Erhebungsverfahren	Telefonisch		Telefonisch		Telefonisch		Face-to-Face		Face-to-Face	
Stichprobenverfahren	Gabler-Häder		Gabler-Häder		Gabler-Häder		Einwohnermeldeamt		Einwohnermeldeamt	
Zustande gekommene Kontakte	**2.213**	**100,0%**	**2.869**	**100,0%**	**4.846**	**100,0%**	**~ 4.119**	**100,0%**	**~2.023**	**100,0%**
Sprachprobleme	144	6,5%	133	4,6%	172	3,5%	~ 65	1,6%	~2	0,1%
Zielperson zu alt	39	1,8%	98	3,4%	113	2,3%	~ 65	4,0%	~37	1,6%
Zielperson krank	42	1,9%	36	1,3%	30	0,6%				
Zielperson in Feldzeit nicht erreichbar	43	1,9%	89	3,1%	75	1,5%	~ 91	2,2%	~21	1,0%
keine Zeit	96	4,3%	264	9,2%	442	9,1%	~165	4,0%	~33	1,6%
kein Interesse	617	27,9%	739	25,8%	1.480	30,5%				
Interview abgebrochen	21	0,9%	16	0,6%	26	0,5%				
Gatekeeper verweigert Zugang zur Zielperson	15	0,7%	9	0,3%	32	0,7%				
Gatekeeper legt wortlos auf	55	2,5%	508	17,7%	695	14,3%				
Anderer Ausfallgrund	113	5,1%	304	10,6%	531	11,0%	~1.697	41,2%	~829	41,0%
Verzerrende Ausfälle II: Verweigerungen u. a.	**1.185**	**53,5%**	**2.196**	**76,5%**	**3.596**	**74,2%**	**~2.083**	**53,0%**	**~922**	**45,4%**
Ausschöpfungsquote (realisierte Interviews)	**1.028**	**46,5%**	**673**	**23,5%**	**1.250**	**25,8%**	**2.036**	**49,4%**	**1.102**	**54,5%**
Quelle	Otte 2002: 88-90		Lamnek / Baur		Arnold		Schneekloth / Leven 2003: 23-34		Schneekloth / Leven 2003: 23-34	

- Mangelnde Befragungsfähigkeit I (Sprachprobleme): Aussiedler, Migranten und Ausländer der zweiten und dritten Einwanderergeneration können sich häufig an Studien nicht beteiligen, weil sie die deutsche Sprache nicht ausreichend beherrschen (Schneekloth / Leven 2003: 30; Otte 2002: 94). Beim ALLBUS kam dieser Ausfallgrund sehr selten vor, was nicht weiter verwunderlich ist, da eine Einwohnermeldeamtstichprobe der wahlberechtigten deutschen Bevölkerung gezogen wurde. Die betroffenen Personengruppen sind aber selten wahlberechtigt. Wie Tabelle 3 zeigt, scheitern je nach Telefonumfrage zwischen 1,5% und 6,5% der Kontakte an der mangelnden Verständigungsmöglichkeit zwischen Interviewer und Zielperson. Dabei ist kein zeitlicher Trend zu erkennen. Unklar ist, ob diese Schwankungen zufällig zustande kommen oder ob sie z. B. auf unterschiedliches Bemühen der Interviewer zurückzuführen sind.
- *Mangelnde Befragungsfähigkeit II (Krankheit, Gebrechlichkeit, Hörprobleme):* Krankheit und Gebrechlichkeit können dazu führen, dass Menschen nicht gewillt sind oder es ihnen nicht möglich ist, sich an einer Umfrage zu beteiligen. Besonders betroffen sind Taubstumme, Schwerhörige, Sprachbehinderte, geistig Behinderte und Demente, also wiederum eine spezifischen Subpopulation, die mit einer zweiten Subpopulation – den Hochbetagten – überlappt (Schneekloth / Leven 2003: 30; Schräpler 2000: 147; Otte 2002: 94). Hörprobleme, Krankheit und Alter als Ausfallgründe nennen in den hier betrachteten CATI-Umfragen je nach Studie zwischen etwa 1,5% bis 4,5% der Befragten. Wieder ist kein zeitlich klarer Trend erkennbar.

  Ein möglicher Grund für die Schwankungen zwischen Studien ist, dass – je nach Studie und Interviewer – derselbe Ausfallgrund unterschiedlich klassiert wird. Hörprobleme sind z. B. nicht immer von Sprachproblemen, Krankheit und altersbedingter Verweigerung zu unterscheiden. Addiert man die verschiedenen Gründe für die mangelnde Befragungsfähigkeit, ergibt sich, dass etwa 5% bis 10% der erreichten Personen deshalb nicht befragt werden können. Zum Teil befinden sich hierunter sicherlich „versteckte“ Verweigerer, da sowohl bei der Studie von Blasius und Reuband zu Beginn der 1990er als auch beim ostdeutschen ALLBUS 2000 nur etwa 2% der erreichten Haushalte aus diesem Grund ausfielen.
- *Mangelnde Erreichbarkeit der Zielperson*: Selbst wenn der Haushalt erreicht wurde, sind zwischen 1% und 6,5% der Zielpersonen während der Feldzeit nicht erreichbar, wobei auch hier kein zeitliches Muster zu erkennen ist. Wie oben erörtert, kann dieses Problem durch längere Feldphasen und durch Terminvereinbarungen verringert werden. Umgekehrt ist es nicht unbedingt ein Zeichen der Güte, wenn dieser Ausfallgrund selten vorkommt: Werden Interviewer leistungsorientiert nach der Zahl der zustande gekommenen Interviews bezahlt, ist die Versuchung groß, das Screening nicht so genau zu nehmen und

einfach eine andere Person im Haushalt zu befragen. Wichtig ist hier also – neben einer angemessenen Basisbezahlung – auch eine gute Interviewerschulung und Supervision. Dass Letzteres leichter zu bewerkstelligen ist, ist einer der großen Vorteile von CATI-Erhebungen.

- *Zeitknappheit (Häder 1994: 19):* Zwischen 2% und 9% der Befragten verweigern die Teilnahme an der Studie, weil sie während der Feldphase keine Zeit für ein Interview haben. Dabei schwankte dieser Ausfallgrund vor fünf Jahren noch zwischen 2% und 4%, während 2005 9% der Befragten dies als Grund angaben. Auffällig ist, dass das Argument des Zeitmangels bei einer achtwöchigen Feldphase genauso oft genannt wird wie bei einer zweiwöchigen. Ob es sich hierbei um tatsächlichen Zeitmangel infolge stärkerer beruflicher und privater Flexibilisierung handelt oder um die mangelnde Bereitschaft, für eine wissenschaftliche Studie Freizeit zu opfern – also implizit um einen Imageverlust der empirischen Sozialforschung –, ist unklar.
- *Mangelnde Kooperationsbereitschaft I (mangelndes Interesse):* Zielpersonen können weiterhin schlicht nicht bereit sein, mit dem Forscherteam zu kooperieren (Otte 2002: 94), z. B. weil sie sich für Wissenschaft an sich oder das Thema der Studie nicht interessieren. Vermutet wird, dass Angehörige der Mittelschicht, höher Gebildete und an Politik und/oder gesellschaftlichen Themen Interessierte eher an Studien teilnehmen als andere Bevölkerungsgruppen (Schneekloth / Leven 2003: 31; Porst et al. 1998: 4; Häder 1994: 19). Hierzu ist dringend weitere Forschung erforderlich, da je nach Studie zwischen einem Viertel und zwei Fünftel der Befragten ihr Desinteresse an der Teilnahme der Studie bekundeten. Mangelndes Interesse ist damit mit Abstand der häufigste Verweigerungsgrund. Hierbei ist allerdings kein klarer Trend erkennbar. Vielmehr ist zu vermuten, dass sowohl das jeweilige Thema der Studie als auch die gesamte Studienorganisation einen starken Einfluss auf diesen Verweigerungsgrund haben. Dabei kommt dem Interviewer eine Schlüsselrolle dabei zu, die Zielpersonen von der Wichtigkeit der Studie zu überzeugen: Befragte bestätigen, dass der Interviewer bedeutsam war für ihre Entscheidung, an der Studie teilzunehmen (Porst 1998: 11). Bei Face-to-Face-Interviews erzielen junge Frauen mit niedriger Bildung und Interviewer ab 70 höhere Ausschöpfungsquoten und sind bessere Konvertierer (Schräpler 2000: 145-146).

  Bei Telefoninterviews ist die Bedeutung von Sprechtempo, Stimmlage und -qualität nicht zu unterschätzen. Am bedeutsamsten scheint aber die Motivation der Interviewer selbst: Schlecht bezahlte und demotivierte Interviewer und ein hoher Zeitdruck verringern die Teilnahmebereitschaft (Porst et al. 1998: 4). Umgekehrt fallen die vergleichsweise geringen Raten an desinteressierten Zielpersonen bei den zwei Eichstätter Studien auf, obwohl es sich um zwei völlig unterschiedliche Themen handelte. In beiden Fällen wurden ausschließlich studentische Interviewer eingesetzt. Bevorzugt wurden angehende

Soziologen. Bei den Interviewerschulungen wurde viel Zeit darauf verwendet, die *Interviewer* davon zu überzeugen, warum die Studie wichtig sei, und ihnen Argumente an die Hand zu geben, wie sie Verweigerer konvertieren könnten. Interviewer mit einer hohen Verweigerungsrate wurden individuell nachgeschult. Bei der „Hartz IV"-Studie hatten die meisten Interviewer zudem ein Eigeninteresse an einer hohen Datenqualität, da sie in das Gesamtprojekt eingebunden waren und die Daten später in Seminararbeiten verwenden mussten.

- *Mangelnde Kooperationsbereitschaft II (Abbruch des Interviews):* Eng verwandt hiermit ist ein weiteres Methodenproblem: Dauert die Befragung zu lange, werden Itembatterien monoton abgefragt oder sind die Fragen uninteressant, steigt die Abbruchquote (Porst et al. 1998: 4). Bei allen hier betrachteten Befragungen war dieses Problem aber vergleichsweise gering: Nur etwa 1% der Zielpersonen brachen ein einmal begonnenes Interview ab. Ist die Teilnahmebereitschaft erst einmal gewonnen, halten also die meisten Befragten das Interview auch bis zum Ende durch, frei nach dem Motto: „Ganz oder gar nicht".
- *Mangelnde Kooperationsbereitschaft III (fehlende subjektive Kompetenz):* Befragte können auch die Teilnahme an einer Studie verweigern, weil sie sich nicht befähigt fühlen, die Fragen zu beantworten (Häder 1994: 20). Über die Häufigkeit des Auftretens dieses Phänomens geben frühere Veröffentlichungen über Ausfallgründe bei CATI-Umfragen keine Hinweise.
- *Mangelnde Kooperationsbereitschaft IV (Misstrauen):* Befragte verweigern schließlich aus Misstrauen. Zu dieser Gruppe von Ausfallgründen gehören erstens Einwände gegen das Telefon als Kommunikationsmedium (Häder 1994: 20). Zweitens misstrauen insbesondere Personen aus dem intellektuellem bzw. alternativen Milieu Datenerhebungen. Ihnen ist der Schutz ihrer Privatsphäre und der Datenschutz besonders wichtig. Institutionen jeder Art begegnen sie kritisch (Schneekloth / Leven 2003: 30; Porst et al. 1998: 4; Häder 1994: 20). So weiß man, dass Personen, die an Umfragen *teilnehmen*, diese für seriös und nach den Regeln des Datenschutzes durchgeführt erachten (Porst 1998: 11). Drittens können Viktimisierungsängste eine Rolle spielen. So verweigern bei persönlich-mündlichem Befragen überproportional viele Hochhausbewohner und Großstädter. Insbesondere männliche Interviewer haben hier große Verweigerungsraten (Schneekloth / Leven 2003: 30; Porst et al. 1998: 4; Schräpler 2000: 144-145). Darüber, wie groß die Rolle des Misstrauens bezüglich der Teilnahmeverweigerung bei telefonischen Befragungen ist, geben die früheren Studien wenig Aufschluss. Allerdings zeigt sich im Zeitverlauf, dass der Anteil der Gatekeeper, die wortlos auflegen, von um die 5% Ende der 1990er bis auf 14% (Studie „Tageszeitungen") bzw. 18% (Studie „Hartz IV") deutlich gestiegen ist. Auch die „sonstigen Ausfallgründe" sind in den vergangenen fünf Jahren von rund 4% bis 7% auf knapp 11% gestiegen.

Insgesamt sind also die Verweigerungsraten in den vergangenen fünf Jahren gestiegen, v. a. durch den Anstieg der wortlos Auflegenden und der sonstigen Ausfallgründe (wobei – wie bereits erwähnt – diese Trendaussage wegen der geringen Zahl der miteinander verglichenen Studien nur unter Vorbehalt gemacht werden kann). Um Anhaltspunkte für die Ursachen des Anstiegs der Verweigerungsraten zu bekommen, betrachte ich abschließend die Ausfallgründe der Studien „Tageszeitungen" und „Hartz IV" aus dem Jahr 2005 differenziert (vgl. Tabelle 4). Die Daten sollten allerdings nur als grobe Tendenzen interpretiert werden, da nicht sichergestellt werden kann, dass alle Interviewer (trotz Schulung) nicht eindeutige Fälle in der relativ langen Liste möglicher Ausfallgründe gleich einordneten.

- *Mangelnde Befragungsfähigkeit:* In konstant etwa 3,5% der Fälle konnten sich die Interviewer mit dem Gatekeeper bzw. der Zielperson nicht verständigen, weil dieser nur sehr schlecht Deutsch sprach und/oder verstand. 2,3% („Tageszeitungen") bzw. 3,5% („Hartz IV") der erreichten Zielpersonen argumentierten, sie seien zu alt, um an der Studie teilzunehmen. Gemäß den Berichten der Interviewer handelte es sich hierbei häufig um eine Mischung aus mangelndem Interesse und mangelnder Befragungsfähigkeit. Möglicherweise ist dies auch der Grund dafür, dass dieser Verweigerungsgrund bei der zweiten Studie seltener auftrat, weil wir hier weitgehend dieselben Interviewer einsetzten und diese mittlerweile erfahrener darin waren, Befragte zu motivieren. Schwerhörigkeit und längere Erkrankungen spielen dagegen eher eine geringe Rolle.
- *Nicht-Erreichbarkeit der Zielperson während der Feldzeit:* Die Gründe, warum Zielpersonen während der Feldzeit nicht erreichbar waren, waren unterschiedlich, und jeder einzelne Grund schlägt für sich quantitativ relativ gering zu Buche. Wie zu erwarten waren bei der „Hartz IV"-Studie die meisten nicht erreichbaren Zielpersonen im Urlaub (Osterferien).
- *Zeitknappheit:* Etwa 3% bis 3,5% der erreichten Zielpersonen meinten, sie hätten keine Zeit, ohne dies weiter zu begründen. Etwas über 1% hatte aufgrund von Freizeitmobilität keine Zeit, z. B. weil sie mit Freunden unterwegs waren oder Verabredungen hatten. Um die 2% der Befragten hatte während der gesamten Feldzeit aus familiären Gründen keine Zeit, z. B. weil sie Familienangehörige pflegen oder Kinder beaufsichtigen mussten. Bei 2,5% der Zielpersonen waren berufliche Anforderungen das Problem. Insgesamt spricht dies für die These, dass in den vergangenen Jahren die Zeitknappheit aufgrund steigender Flexibilisierungsanforderungen zugenommen hat und dass dies auch die Teilnahmebereitschaft bei CATI-Umfragen beeinflusst.
- *Kein Interesse:* Etwa 5% der erreichten Personen gaben offen zu, keine Lust zu haben, Freizeit für eine wissenschaftliche Studie zu opfern. Weitere 20 bzw. 25% verweigerten allgemein aus Desinteresse.

- *Gatekeeper blockt ab:* Deutlich weniger als 1% der Gatekeeper verweigerte den Zugang zur Zielperson. In einigen wenigen Fällen war ein Kleinkind am Telefon, das das Telefon nicht an einen Erwachsenen weitergab. Dieses Problem ließ sich aber durch einen erneuten Kontaktversuch leicht lösen.
- *Auflegen:* Wortlos aufzulegen – die Kommunikation also zu unterbrechen, ohne dem Gegenüber eine Möglichkeit zu geben, einen vom Gegenteil zu überzeugen – ist ein deutliches Zeichen von Misstrauen gegenüber in Marketing- und Überzeugungstechniken geschulten und damit rhetorisch überlegenen Fremden. Da der Anteil der Befragten, die wortlos auflegten, seit 2001 so stark gestiegen ist, erfassten wir den *Zeitpunkt*, zu dem aufgelegt wurde:
  - *Sofortiges Auflegen:* 5,4% (Studie „Tageszeitungen") bzw. 6,8% (Studie „Hartz IV") der Gatekeeper legten sofort auf, sobald sie hörten, dass eine fremde Person am Telefon war, noch bevor der Interviewer den zweiten Satz beenden konnte. Ich vermute, dies ist auf die drastische Zunahme der (teils als Umfragen getarnten) Telefon-Marketing-Anrufe in den vergangenen zwei Jahren zurückzuführen (test 2005). Deshalb kommt dem Begrüßungssatz – mehr denn je – eine entscheidende Bedeutung zu. Bei beiden Studien lautete dieser: „Guten Tag, hier ist die Universität Eichstätt-Ingolstadt Mein Name ist (...)." Diese Anrede wurde nach dem (bezüglich des sofortigen Auflegens noch katastrophaler verlaufenden) Pretest gewählt, um sofort deutlich zu machen, dass der Anrufer einer seriösen wissenschaftlichen Einrichtung angehörte.
  - *Auflegen während der Begrüßungsfrage:* Nachdem die Interviewer ihren Namen genannt hatten, stellten sie die Begrüßungsfrage. Diese lautete bei der Studie „Tageszeitungen": „Wir führen eine wissenschaftliche Studie zum Thema Tageszeitungen durch. Dazu würden wir gerne eine Person aus ihrem Haushalt befragen. Das Interview dauert etwa 15 Minuten. Wir bitten herzlich um ihre Unterstützung." Bei der „Hartz IV"-Studie sagte der Interviewer: „Wir führen eine wissenschaftliche Studie zu den Themen Arbeitslosigkeit und Sozialstaat durch. Dazu würden wir gerne eine Person aus ihrem Haushalt befragen. Das Interview dauert etwa 15 Minuten. Wir bitten herzlich um ihre Unterstützung." Auf Anfrage erläuterten die Interviewer die Stichprobenstrategie, die Maßnahmen zur Gewährleistung des Datenschutzes, die Finanzierung bzw. die Studienziele, und sie gaben Telefonnummer, Email- und Webadresse der Studienleitung bekannt. Rund 8% (Studie „Tageszeitungen") bzw. 10% (Studie „Hartz IV") der Gatekeeper legte spätestens jetzt wortlos auf. Beim Pretest der ersten Studie lagen diese Raten noch höher. Dies unterstreicht, wie zentral das Verhalten des Interviewers in diesen ersten ein bis zwei Minuten ist, aber auch, dass das Misstrauen gegenüber Befragungen jeder Art gestiegen ist.

  - *Auflegen während des Screenings / des Hauptteils:* Bestärkt wird die Hypothese des gestiegenen Misstrauens dadurch, dass während des Screenings (5 Fragen bei „Hartz IV“, 4 Fragen bei „Tageszeitungen“) und während des Hauptteils (der in beiden Fällen im Schnitt deutlich länger als 15 Minuten dauerte) jeweils weniger als 1% der Kontaktierten das Gespräch beendete.
- *Offen bekundetes Misstrauen:* Ein weiteres Indiz für das gestiegene Misstrauen gegenüber Telefoninterviews sind die Gründe, die für mangelnde Kooperationsbereitschaft angegeben werden: Bei beiden Studien meinten etwa 2,5% der Befragten, dass sie keine Fragen am Telefon beantworteten. Weitere 2,2% beteiligen sich grundsätzlich bei keinen Umfragen. Zweifel bezüglich des Datenschutzes hegte dagegen nur eine Minderheit von 0,1% bzw. 0,2% der Befragten. Angesichts der Tatsache, dass etwa ein Drittel der Haushalte nicht registrierte Nummern haben, ist es auch interessant zu vermerken, dass nur 0,2% bzw. 0,3% der erreichten Haushalte dies als Verweigerungsgrund angaben.
- *Sonstige Ausfallgründe waren relativ unbedeutend:* Weniger als ein halbes Prozent der Zielpersonen wagte es auch nach gutem Zureden seitens des Interviewers nicht, an der Umfrage teilzunehmen, weil sie sich angeblich nicht auskannten und deshalb für inkompetent hielten. Auch die bisweilen geäußerte Befürchtung, Incentives würden „den Markt verderben“, kann hier nicht bestätigt werden: Nur 0,1% der Befragten verlangten eine Bezahlung für eine Beteiligung an einer Studie. 0,6% bzw. 0,7% der Befragten gaben an, sie seien erst vor kurzem befragt worden. Bei „Hartz IV“ konnten wegen der kurzen Feldphase nicht alle Terminvereinbarungen abgearbeitet werden, weil einige Personen zum vereinbarten Termin nicht anwesend waren – wahrscheinlich handelt es sich hierbei um versteckte Verweigerungen.

Tabelle 4: Verweigerungsrate differenziert nach Verweigerungsgründen

Erhebungszeitraum	**März 2005**		**April – Juni 2005**	
Region	**Bremen, Baden-Württemberg, NRW, Sachsen-Anhalt**		**BRD**	
Studie	**Hartz IV**		**Qualität von Tageszeitungen**	
Erhebungsverfahren	**Telefonisch**		**Telefonisch**	
Stichprobenverfahren	**Gabler-Häder**		**Gabler-Häder**	
Zustande gekommene Kontakte	**2.869**	**100,0%**	**4.846**	**100,0%**
Sprachprobleme	101	**3,5%**	172	**3,5%**
krank	36	1,3%	30	0,6%
Hörgeschädigter	32	1,1%	32	0,7%
Alter	98	**3,4%**	113	**2,3%**
Urlaub	20	0,7%	20	0,4%
Geschäftsreise	2	0,1%	3	0,1%
Abwesend aus familiären Gründen	1	< 0,1%		
In Feldzeit nicht erreichbar (Sonstiges)	66	2,3%	52	1,1%
keine Zeit (Beruf)	72	**2,5%**	132	**2,7%**
keine Zeit (Familie)	65	**2,3%**	84	**1,7%**
keine Zeit (Freizeit)	41	**1,4%**	59	**1,2%**
keine Zeit (Sonstiges)	86	**3,0%**	167	**3,4%**
kein Interesse	584	**20,4%**	1.254	**25,9%**
keine Lust. Freizeit zu opfern	155	**5,4%**	226	**4,7%**
Gatekeeper blockt ab	9	0,3%	30	0,6%
Kind			2	< 0,1%
Auflegen (sofort)	194	**6,8%**	263	**5,4%**
Auflegen (Begrüßungsfrage)	293	**10,2%**	399	**8,2%**
Auflegen (Screening)	21	0,7%	33	0,7%
Auflegen (Hauptteil)	16	0,6%	26	0,5%
beantwortet keine Fragen am Telefon	73	**2,5%**	118	**2,4%**
macht nicht bei Umfragen mit	62	**2,2%**	109	**2,2%**
Datenschutz	7	0,2%	4	0,1%
Geheimnummer	7	0,2%	16	0,3%

kennt sich nicht aus	11	0,4%	11	0,2%
Bezahlung	2	0,1%	5	0,1%
wurde vor kurzem befragt	16	0,6%	32	0,7%
Bundesland nicht genannt	1	< 0,1%		
Terminvereinbarung	43	1,5%	20	0,4%
Sonstiges	82	2,9%	184	3,8%
Verzerrende Ausfälle II: Verweigerungen u. a.	**2.196**	**76,5%**	**3.596**	**74,2%**
Ausschöpfungsquote (realisierte Interviews)	**673**	**23,5%**	**1.250**	**25,8%**
Quelle	Lamnek / Baur		Arnold	

6 Schlussfolgerungen

Noch vor wenigen Jahren schien das Stichprobenproblem bei Telefonumfragen durch das Gabler-Häder-Design gelöst. Insgesamt ist in den vergangenen Jahren Undercoverage aber wieder zu einem Problem geworden: Einerseits steigt die Zahl der nur noch mobil oder per Internet-Telefon erreichbaren Personen (vgl. Gabler /Häder, in diesem Band). Andererseits steigt die Zahl der Verweigerungen. Diese Aussage kann allerdings nur mit Vorbehalt gemacht werden, da bislang nur für sehr wenige Studien die Ausfallgründe veröffentlicht worden sind, so dass es sich hierbei auch um zufällige Schwankungen oder um Institutseffekte handeln kann. Die wenigen verfügbaren Daten legen folgende Vermutungen nahe:

Bereinigt man die Bruttostichprobe um alle eindeutig oder auch nur möglicherweise stichprobenneutralen Ausfälle, so sanken Ausschöpfungsquoten mit dem Gabler-Häder-Design von rund 36% bis 47% um die Jahrtausendwende auf 23% bis 26% im Jahr 2005.

Der seit Mitte der 1990er mit Abstand wichtigste Ausfallgrund war mangelndes Interesse an der Studie. Die Ausfallrate aus diesem Grund blieb allerdings relativ konstant. Der Anstieg der Verweigerungsrate ist vielmehr auf zunehmende Zeitknappheit (vor allem aus beruflichen und familiären Gründen) und insbesondere auf gestiegenes Misstrauen gegenüber Befragungen zurückzuführen. Etwa 14% bis 17% der Gatekeeper legen das Telefon kommentarlos auf, noch bevor der Interviewer sein Anliegen vorgestellt hat. Etwa weitere 5% der Gatekeeper äußern offen ihr Misstrauen gegenüber Befragungen und weigern sich deshalb teilzunehmen. Als Ursache vermute ich die massive Zunahme des Telefonmarketing mit Autodialern insbesondere in den letzten drei Jahren (test 2005).

Da der Einsatz der Inferenzstatistik und damit der größte Vorteil der quantitativen gegenüber der qualitativen Sozialforschung von Zufallsstichproben abhängt,

ist das Problem von durch so massive Ausfälle möglicherweise auftretenden Verzerrungen nicht zu unterschätzen. Längst bekannte Maßnahmen zur Erhöhung der Ausschöpfungsquote werden umso wichtiger:

- *Ein gut formulierter Gesprächseinstieg* (Porst et al. 1998: 13-14; 21) wird immer zentraler, da die Entscheidung, den Hörer aufzulegen, häufig bereits innerhalb der ersten Sekunden fällt.
- *Qualifizierte Interviewer spielen* bei der Umsetzung dieses Einstiegs eine große Rolle. Wichtig sind „eine qualifizierte Schulung der Interviewer, daraus resultierend ein qualifiziertes Interviewerverhalten vor und während der Befragung; ein bestimmter Interviewertyp (‚little old ladies'); erfahrene Interviewer mit positiven Erwartungen an ihre Arbeit und mit dem Glauben an die eigenen Fähigkeiten; flexible Interviewer, die sich vor allem während der Kontaktphase den unterschiedlichsten Zielpersonen angemessen präsentieren" (Porst et al. 1998: 20-21).
- *Warmkontakte* (Porst et al. 1998: 11-13; 21; Hüfken 2000): „Ein kurzes Anschreiben kann ganz generell als ‚Schlüssel' zum Telefoninterview gesehen werden (...); der Forscher kann darlegen, warum es wichtig ist, an der angekündigten Befragung teilzunehmen, er kann die Vertraulichkeit der Befragung betonen und versuchen, Vertrauen in die durchführende Einrichtung zu wecken" (Porst et al. 1998: 12). Das Problem hierbei ist, dass Stichproben mit dem Gabler-Häder-Design per Definition Kaltkontakte sind, da dem Institut nur die Telefonnummer, sonst aber keine Informationen über die Befragten vorliegen. Meines Wissens erlauben Telefon-CDs nur, die zu einer Adresse gehörende Telefonnummer zu recherchieren, aber nicht, die zu einer Telefonnummer gehörende Adresse. Damit ist bedauerlicherweise ein noch vor kurzem extrem leistungsfähiges und kostengünstiges Instrument der quantitativen Sozialforschung nur noch für solche Studien brauchbar, bei denen kein Zusammenhang zwischen den hier genannten Ausfallgründen und Studienziel zu vermuten ist. Vielmehr würde ich empfehlen, zumindest für wissenschaftliche Studien künftig wieder auf Einwohnermeldeamtsstichproben zurückzugreifen.

Gleichzeitig wird aber deutlich, dass wir immer noch zu wenig über Verweigerer wissen. Um diese Lücken zu schließen, sind mindestens drei Maßnahmen erforderlich:

- *Detaillierte Ausfallprotokolle:* Für jede Telefonumfrage sollte gemäß Schnells (1997) Vorschlag ein detailliertes Ausfallprotokoll geführt werden. Bei den CATI-Umfragen ist dies i. d. R. kein Problem, da die meisten CATI-Programme dies unterstützen. Erst wenn über eine genügend große Zahl von Studien Ausfallstatistiken vorliegen, kann entschieden werden, ob es sich bei den hier

präsentierten Daten um einen wirklichen Trend, um Institutseffekte oder um zufällige Schwankungen handelt.

- *Studien über Maßnahmen zur Erhöhung der Ausschöpfungsquoten:* Mittels sozialwissenschaftlicher Experimente – in diesem Fall also eine systematische Variation von Studieninhalten, Instituten, Schulungsleitern, Supervisoren, Interviewern, Bezahlungssystem usw. – könnte quantifiziert werden, welchen Einfluss die genannten Faktoren auf die Ausschöpfungsquoten haben.
- *Studien über die Differenzen zwischen Verweigerern und Studienteilnehmern:* Weiterhin sind dringend Studien über Einstellungen und Verhaltensweisen der verschiedenen Subpopulationen der Verweigerer erforderlich. Da sich diese Personenkreise systematisch standardisierten Erhebungen entziehen, sehe ich die einzige Möglichkeit hierzu in qualitativen Studien über diese Gruppe. Hierbei wären völlig andere Zugangswege zu wählen, z. B. im Rahmen eines wissenschaftlichen Projekts über das Schneeballprinzip, in der Hoffnung, dass zumindest ein Teil der Verweigerer keine Einwände gegen sozialwissenschaftliche Studien an sich hat, sondern nur gegen kommerzielle Sozialforschung und/oder standardisierte Befragungen.

7 Literatur

Behnke, Joachim; Baur, Nina & Behnke, Nathalie (2006): Empirische Methoden der Politikwissenschaft. Paderborn: Schöningh

Blasius, Jörg & Reuband, Karl-Heinz (1995): Telefoninterviews in der empirischen Sozialforschung: Ausschöpfungsquoten und Antwortqualität. In: ZA-Informationen. Heft 37. S. 64-87. Zu lesen: S. 64 und S. 84-85

Deutschmann, Marc & Häder, Sabine (2002): Nicht-Eingetragene in CATI-Surveys. In: Gabler, Siegfried & Häder, Sabine (Hg.) (2002); Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. München / Berlin: Waxmann Münster / New York. 68-84

Gabler, Siegfried & Häder, Sabine (1997): Überlegungen zu einem Telefonstichprobendesign für Telefonumfragen in Deutschland. In: ZUMA Nachrichten. Band 21. Heft 41. 7-18

Gabler, Siegfried & Häder, Sabine (1998): Probleme bei der Anwendung von RLD-Verfahren. In: Gabler, Siegfried, Häder, Sabine & Hoffmeyer-Zlotnik, Jürgen H.P. (Hg.) (1998): Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag. 58-68

Gabler, Siegfried & Häder, Sabine (1999): Erfahrungen beim Aufbau eines Auswahlrahmes für Telefonstichproben in Deutschland. In: ZUMA Nachrichten. Band 23. Heft 44. 45-61

Gabler, Siegfried & Schürle, Josef (2002): Zur Stabilität des Gabler-Häder-Auswahlrahmens. In: Gabler, Siegfried & Häder, Sabine (Hg.) (2002); Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. München / Berlin: Waxmann Münster / New York. 59-67

Häder, Sabine (1994): Auswahlverfahren bei Telefonumfragen. ZUMA-Arbeitsbericht 94/03. http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte am 09.12.2004

Häder, Sabine (1996): Wer sind die „Nonpubs"? Zum Problem anonymer Anschlüsse bei Telefonumfragen. In: ZUMA-Nachrichten. Band 20. Heft 39. 45-68

Häder, Sabine & Gabler, Siegfried (1998): Ein neues Stichprobendesign für telefonische Umfragen in Deutschland. In: Gabler, Siegfried; Häder, Sabine & Hoffmeyer-Zlotnik, Jürgen H. P. (Hg.) (1998): Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag. 69-88

Häder, Sabine & Gabler, Siegfried (2000): Überlegungen zur Anwendung von RLD-Verfahren bei Telefonumfragen in Deutschland. In: Hüfken, Volker (Hg.) (2000): Methoden in Telefonumfragen. Wiesbaden: Westdeutscher Verlag. 33-47

Häder, Sabine (2000): Telefonstichproben. ZUMA How-to-Reihe. Nr. 6. Mannheim: ZUMA. http://www.gesis.org/Publikationen/Berichte/ZUMA_How_to/Dokumente/pdf/how-to6sh.pdf am 09.12.2004-12-09

Hüfken, Volker (2000): Kontaktierung bei Telefonumfragen. Auswirkungen auf Kooperations- und Antwortverhalten. In: Hüfken, Volker (Hg.) (2000): Methoden in Telefonumfragen. Wiesbaden: Westdeutscher Verlag. 11-33

Otte, Gunnar (2002): Erfahrungen mit zufallsgenerierten Telefonstichproben in drei lokalen Umfragen. In: Gabler, Siegfried & Häder, Sabine (Hg.) (2002); Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. München / Berlin: Waxmann Münster / New York. 85-110

Porst, Rolf (1998): Erfahrungen mit und Bewertung von Umfragen. Was unsere Befragten über Umfragen denken. ZUMA-Arbeitsbericht 98/03. http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte am 09.12.2004

Porst, Rolf, Ranft, Sabine & Ruoff, Bernd (1998): Strategien und Maßnahmen zur Erhöhung der Ausschöpfungsraten bei sozialwissenschaftlichen Umfragen. Ein Literaturbericht. ZUMA-Arbeitsbericht 98/07. http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte/documents/pdfs/98_07.pdf am 09.12.2004

Schneekloth, Ulrich & Leven, Ingo (2003): Woran bemisst sich eine „gute" Allgemeine Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. In: ZUMA-Nachrichten. Band 27. Heft 53. 16-57

Schnell, Rainer (1997): Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Opladen: Leske + Budrich.

Schräpler, Jörg-Peter (2000): Was kann man am Beispiel des SOEP bezüglich Nonresponse lernen? In: ZUMA-Nachrichten. Band 24. Heft 46. 117-150

Stögbauer, Andrea (2000): Ausschöpfungsprobleme telefonischer Umfragen. Eine Zwischenbilanz praktischer gesamtdeutscher Erfahrung. In: Hüfken, Volker (Hg.) (2000): Methoden in Telefonumfragen. Wiesbaden: Westdeutscher Verlag. 91-104

test (2005): Briefe mit Blutspur. In: test 11/2005. 14-17

Anreizeffekte in Studien der Markt- und Sozialforschung

Uwe Engel

1 Einführung

Eine für die Markt- und Sozialforschung bedeutsame Frage ist darin zu sehen, inwieweit durch das Setzen von Anreizeffekten der Ausschöpfungsgrad ihrer Studien erhöht werden kann. Für den Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM) haben wir daher eine Metaanalyse des Forschungsstandes durchgeführt und deren Ergebnisse in einem Forschungsbericht auf der Homepage des ADM publiziert (http://www.adm-ev.de, Rubrik „Forschungsprojekte“)[1].

Vorliegender Beitrag stützt sich auf eines der dort beschriebenen Modelle und erweitert es um eine Analyse des Effektes der Fragebogen- bzw. Interviewlänge auf die Stärke des Anreizeffektes auf den Ausschöpfungsgrad. Darüber hinaus wird Evidenz aus zwei weiteren Studien in die Betrachtung einbezogen, und zwar aus einem Studienbarometerprojekt, das wir an der Universität Bremen durchgeführt haben, sowie aus einer Analyse mit Daten des „Sozioland“-Access Panels.

2 Metaanalyse

Die Metaanalyse, die wir für den ADM zur Stärke von Anreizeffekten durchgeführt haben, stützt sich auf berichtete Antwortraten aus 68 Studien, die es erlaubten, Anreizeffekte unter zusammengenommen 177 Studienbedingungen zu untersuchen. Zu diesen Bedingungen zählten als erstes die unmittelbaren Anreizbedingungen selbst. Dazu gehörte, ob es sich um monetäre oder nicht-monetäre Anreize handelte, ob sie im Voraus („prepaid“) gewährt oder für den Fall in Aussicht gestellt („promised“) wurden, dass z.B. der Fragebogen ausgefüllt retourniert wurde. Des Weiteren konnte berücksichtigt werden, ob ein – dann in aller Regel in Aussicht gestellter – Anreiz eine sichere Belohnung darstellte oder lediglich die Teilnahme an einer Lotterie vorsah, infolge derer man nur mit einer bestimmten Wahrscheinlichkeit in den Genuss einer Belohnung oder Gegenleistung kom-

1 Für Anlage, Details und Ergebnisse der Metaanalyse wird der interessierte Leser auf die oben bezeichnete Quelle und die Möglichkeit verwiesen, sich den Forschungsbericht dort herunterzuladen. Dem ADM danken wir für die finanzielle Unterstützung des Projekts.

men konnte. Schließlich konnte berücksichtigt werden, ob die zu erhaltende Belohnung für einen gemeinnützigen Zweck gespendet oder dem Respondenten selbst zugute kommen sollte. Die Metaanalyse wurde entsprechend so aufgebaut, dass sie den Anreizeffekt als Funktion der Unterschiede monetär vs. nicht monetär, „prepaid" vs. „promised" (sicher), Lotterieteilnahme vs. sichere Belohnung sowie Belohnung für gemeinnützigen Zweck vs. für den Respondenten selbst untersuchte. Außerdem konnte die Wirkung der Anreiz- bzw. Geldhöhe in die Analyse einbezogen werden.

Das Modell, das die Effekte dieser unmittelbaren Anreizbedingungen schätzte, haben wir vorliegendem Beitrag zugrunde gelegt (Tabelle 1). Auf die darüber hinaus gerechneten Modelle soll hier nur verwiesen werden. Sie kontrollieren zusätzlich a) die Effekte der Höhe der Antwortrate, b) die Art des Samples/der Population (Bevölkerungsumfrage, Konsumentenbefragung, spezielle Berufsgruppe, sonstige spezielle Population), c) zwei Ländergruppen, d) den Zeitraum, in den die Studie fiel (1970er, 1980er, 1990er) und e) die Erhebungsmethode inkl. Feldzugangsmodalitäten i.S. von Dillman's „Tailored Design-" Methode.

3 Methode

Zur Analyse der Effekte, die auf den Einsatz von Anreizen (Incentives) zurückzuführen sind, berechneten wir für jede Studienbedingung eines der beiden in Metaanalysen üblichen Maße, und zwar das Effektstärkenmaß[2], *d*. Es ist definiert als standardisierte Differenz in den mittleren Werten von Experimental- und Kontrollgruppe, das sind im vorliegenden Zusammenhang die Gruppen mit [„Inc"] versus *ohne* [„Ninc"] Einsatz von Incentives:

$$d = \frac{\bar{y}_{Inc} - \bar{y}_{Ninc}}{s}$$

Die Standardisierung erfolgt dabei so, dass die Mittelwerts- bzw. Anteilswertdifferenz in Standardabweichungseinheiten *s*, also in Einheiten der durchschnittlichen Streuung des zugrunde liegenden Merkmals, ausgedrückt wird.

Dieses *d* (und somit nicht die Höhe der Antwortrate selbst, sondern die Stärke des Anreizeffektes auf diese Rate) stellte die abhängige Größe von Mehrebenenmodellen dar, die die Anreiz- bzw. Studienbedingungen als Determinanten enthält (wobei einzelne Variablenblöcke jeweils schrittweise ergänzt wurden). Die jeweilige Schätzgleichung [3] resultierte aus Einsetzen der Level-2 Gl. [2] in die

2 Vgl. dazu z.B. Hunter, John E.; Frank L. Schmidt (1990) Methods of Meta-Analysis. Newbury Park: Sage, p. 271

Level-1 Gl. [1] und umfasste im Allgemeinen die fixen Effekte b_k der K über die Bedingungen i von Studie j variierenden Merkmale x_k, die fixen Effekte c_l der L über die Studien j variierenden Merkmale v_l sowie die Zufallseffekte auf den beiden Analyseebenen, u_{0j} und e_{ij}. Die b's und c's sind wie gewöhnliche multiple Regressionskoeffizienten im Sinne erwarteter Unterschiede bzw. erwarteter Veränderungen interpretierbar. Im vorliegenden Beitrag sind von den in der Analyse insgesamt berücksichtigten Variablenblöcken nur die unmittelbaren Anreizbedingungen einbezogen (Tabelle 1), ergänzt um die Wirkung der Fragebogen- bzw. Interviewlänge (Tabelle 2).

Modell

$$[1] \qquad d_{ij} = b_{0j} + b_1 x_{1ij} + \sum_{k=2}^{K} b_k x_{kij} + e_{ij}$$

$$[2] \qquad b_{0j} = b_0 + \sum_{l=1}^{L} c_l v_{lj} + u_{0j}$$

$$[3] \qquad d_{ij} = b_0 + b_1 x_{1ij} + \sum_{k=2}^{K} b_k x_{kij} + \sum_{l=1}^{L} c_l v_{lj} + u_{0j} + e_{ij}$$

Auf die darüber hinaus durchgeführten Variationen in der Methode kann hier nur verwiesen werden. Diese betreffen zum einen die Frage, ob die Varianz von d als bekannte oder zu schätzende Größe in die Berechnung eingeht, und zum anderen einen Vergleich in der Modellbildung selbst, der statt des mehrebenenanalytischen Ansatzes den seinerzeit von Coleman entwickelten WLS-Ansatz als Alternative für Zwecke vorliegender Metaanalyse nutzt.

Tabelle 1: Die Stärke des Anreizeffektes als Funktion unmittelbarer Anreizbedingungen

Mit follow-up's	**Modell 1**		**Modell 2**		**Modell 3**		**Modell 4**		**Modell 5**	
n_{1b} = 177; se standard error	b	b/se	b	b/se	b	b/se	b	b/se	b	b/se
Konstante	0,22	13,1	0,23	12,4	-0,021	-0,5	-0,002	-0,0	-0,063	-1,2
Follow-up			-0,024	-1,3	-0,052	-3,3	-0,052	-3,4	-0,055	-4,0
Anreizbedingungen										
Monetär [1] vs. nicht-monetär [0]					0,135	4,0	0,139	4,2	0,125	3,9
Prepaid [1] vs. promised (sicher) [0]					0,221	7,4	0,199	6,1	0,198	6,6
Promised (Lotterie) vs. promised (sicher) [0]					0,039	0,8	0,014	0,3	0,017	0,4
Gemeinwohlkomponente: ja [1] vs. nein [0]							-0,097	-1,6	-0,091	-1,6
Anreizhöhe: nicht variiert [1] vs. Anreizhöhe: gering [0]									0,066	2,0
Mittlere Anreizhöhe [1] vs. Anreizhöhe: gering [0]									0,131	4,0
Höhere Anreizhöhe [1] vs. Anreizhöhe: gering [0]									0,223	5,8
Erklärte Varianz*			0,6%		27,8%		29,3%		36,4%	

*Modelle 3 – 5 (Bezug: Modell 2); Modell 2 (Bezug: Modell 1)

4 Ergebnisse

Tabelle 1 zeigt, welche Faktoren in die Analyse einbezogen wurden und wie sich deren Einfluss auf die Stärke des Anreizeffektes darstellt.

Die durchschnittliche standardisierte Effektstärke über alle Studien und Bedingungen hinweg liegt bei 0,22 bzw. 22% der durchschnittlichen Streuung des zugrunde liegenden Merkmals (Modell 1). Der Wert ist positiv und besagt somit, dass die Response Rate in der Incentives-Gruppe im Allgemeinen über derjenigen der Nicht-Incentives Gruppe liegt. Im Durchschnitt beträgt die Differenz in den Antwortraten beider Gruppen 0,22 Standardabweichungseinheiten.

Dieser Wert verändert sich leicht auf 0,23 für den jeweiligen Erstkontakt, also unter Ausschluss der Anreizeffekte im Kontext von follow-up's. Der für diese follow-up's in Modell 2 ausgewiesene negative Effekt besagt, dass sich die Stärke des Anreizeffektes über die follow-up's hin abschwächt.

Mit Modell 3 nehmen wir ergänzend die erste Gruppe von Anreizbedingungen in die Gleichung auf (monetär vs. nicht-monetär und prepaid vs. promised). Außerdem wird beachtet, dass es sich bei um in Aussicht gestellten Gratifikationen um Lotteriegewinne handeln kann. Die ausgewiesenen Koeffizienten beziehen sich jeweils auf den links außen in der Tabelle spezifizierten Vergleich bzw. Kontrast und liefern für diesen den erwarteten Unterschied in den standardisierten Anreizeffektstärken. Ein positives Vorzeichen des Koeffizienten indiziert dabei, dass die Anreizeffektstärke unter der mit [1] kodierten Bedingung höher ist als unter der mit [0] kodierten Bedingung, und vice versa. So besagt etwa der für den Vergleich „monetär [1] vs. nicht-monetär [0]" ausgewiesene Koeffizient von 0,135, dass bei Einsatz von monetären Anreizen die standardisierte Anreizeffektstärke im Schnitt um 0,135 Einheiten über der standardisierten Anreizeffektstärke bei Einsatz von nicht-monetären Anreizen liegt. Der Einsatz von monetären Anreizen ist also effektiver als der Einsatz nicht-monetärer Anreize, da mit ihnen ein stärkerer Effekt auf die Response Rate erzielt werden kann. Der Unterschied ist bedeutsam, was auch daran erkennbar ist, dass er das Vierfache seines geschätzten Standardfehlers ausmacht. Als noch stärker erweist sich allerdings der Unterschied im Vergleich der Bedingungen „prepaid vs. promised (sicher)". Wir registrieren einen Wert von 0,221, der annähernd dem 7 ½ fachen seines geschätzten Standardfehlers entspricht. Besonders effektiv ist es also, den Anreiz im Voraus zu setzen und nicht davon abhängig zu machen, dass der Respondent zunächst dem Befragungsanliegen entspricht. Letztes ist auch dann nicht effektiv, wenn es in die Form einer Lotterieteilnahme gebracht wird, da wir auch mit dem leichten Zuwachs gegenüber einer in Aussicht gestellten sicheren Gratifikation weit unter der mit der Bedingung „prepaid" erzielbaren Anreizeffektstärke bleiben. Beide „promised" – Varianten stellen keine günstigen Anreizbedingungen dar.

Als tendenziell kontraproduktiv erscheint im Spiegel des Befundes der Einsatz einer Gemeinwohl- oder Altruismuskomponente, die darin besteht, dass die Gratifikation einem gemeinnützigen Zweck und nicht dem Respondenten selbst zukommen soll.

Nach dem Befund in Tabelle 1 ist des Weiteren zu erwarten, dass die Stärke des Anreizeffektes mit der Anreizhöhe steigt. Wir registrieren zwei positive und auch starke Koeffizienten, wenn wir mittlere und höhere Anreizhöhen mit geringer Anreizhöhe in Beziehung setzen (0,13 zu 0,22).

Tabelle 2: Der Einfluss der Fragebogen-/Interviewlänge auf die Stärke des Anreizeffektes

Mit follow-up's	**Modell 6**		**Modell 7**	
n_{1b} = 177; se standard error	b	b/se	b	b/se
Konstante	-0,035	-0,6	0,294	8,2
Follow-up	-0,056	-4,0	-0,028	-1,6
Fragebogen-/Interviewlänge „k.A." [1] vs. „kurz" [0]	-0,017	-0,4	-0,072	-1,7
Fragebogen-/Interviewlänge „mittel" [1] vs. „kurz" [0]	-0,051	-1,0	-0,080	-1,7
Fragebogen-/Interviewlänge „lang" [1] vs. „kurz" [0]	-0,078	-1,3	-0,121	-2,0
Anreizbedingungen				
Monetär [1] vs. nicht-monetär [0]	0,122	3,8		
Prepaid [1] vs. promised (sicher) [0]	0,198	6,6		
Promised (Lotterie) vs. promised (sicher) [0]	0,013	0,3		
Gemeinwohlkomponente: ja [1] vs. nein [0]	-0,092	-1,6		
Anreizhöhe: nicht variiert [1] vs. Anreizhöhe: gering [0]	0,070	1,9		
Mittlere Anreizhöhe [1] vs. Anreizhöhe: gering [0]	0,131	4,0		
Höhere Anreizhöhe [1] vs. Anreizhöhe: gering [0]	0,224	5,8		
Erklärte Varianz*	38,2%		3,3%	

Setzen wir diesen Befund in Beziehung zu den mittleren Beträgen, die in den Studien eingesetzt wurden[3], so wäre eine unmittelbar auf die Anreizhöhe abstellende Interpretation der Anreizeffektstärke allerdings nur unter der Bedingung „promised" (ohne Lotterie), weniger aber für die Bedingung „prepaid" plausibel, wenn als Kriterium jeweils der Vergleich in den mittleren Geld- bzw. Wertbeträgen zugrunde gelegt wird. Eine auf die Anreizhöhe abstellende Interpretation wird auch dadurch infrage gestellt, dass der Befund selbst unter nur leicht veränderten Modellspezifikationen nicht replizierbar ist. So verändert sich das Verhältnis der Koeffizienten für die Kontraste „mittlere vs. geringe" und „höhere vs. geringe" Anreizhöhe von 0,13 zu 0,22 auf 0,16 zu 0,19, wenn die Berechnungen unter Ausschluss der follow-up's durchgeführt werden[4]. Es kann dann kaum noch auf einen auch nur annähernd linearen Effekt der Anreizhöhe auf die Stärke des Anreizeffektes geschlossen werden. Wird zudem eine weitere Bedingung in der Modellspezifikation verändert (Varianz von *d* nicht geschätzt, sondern als bekannte Grö-

3 Vgl. Tabellen 4.4 und 4.5 unseres Forschungsberichts auf http://www.adm-ev.de
4 Vgl. Tabelle 4.9 unseres Forschungsberichts.

ße einbezogen[5]), dann verändert sich besagtes Verhältnis auf 0,17 zu 0,16. Beide Effekte sind dann annähernd gleich und besagen damit, dass sich die Anreizhöhe nicht auf die Anreizstärke auswirkt.

Tabelle 2 informiert über den Einfluss der Fragebogen- bzw. Interviewlänge[6] auf die Stärke des Anreizeffektes. Wie erwartet, sind die Vorzeichen der Effekte negativ. Das bedeutet, dass ein gegebener Anreiz bei einem kurzen Fragebogen bzw. Interview am stärksten wirkt und sich diese Wirkung mit zunehmender Länge von Fragebogen bzw. Interview tendenziell abschwächt.

5 Diskussion

Nach vorliegender Analyse stellt das Setzen von Anreizen ein wirksames Instrument dar. Dabei erweist sich der Einsatz von monetären Anreizen als effektiver als der Einsatz nicht-monetärer Anreize. Entscheidend ist zudem der Zeitpunkt, zu dem ein Anreiz eingesetzt wird. Effektiv sind monetäre Anreize, wenn sie „prepaid" eingesetzt werden, also nicht davon abhängig gemacht werden, dass der Fragebogen erst ausgefüllt zurückgesandt werden muss. Hingegen ist der ausschöpfungserhöhende Effekt von Anreizen, die nur in Aussicht gestellt bzw. versprochen werden, empirisch zweifelhaft. Empirisch zweifelhaft ist auch der Effekt von Anreizen, die in einer Lotterieteilnahme bestehen oder als gemeinnützige Spende gar nicht dem Respondenten selbst zugute kommen sollen.

Warum aber wirken Anreize? Dass insbesondere im Voraus gewährte monetäre Anreize wirken, spricht eher für eine austauschtheoretische als für eine rationaltheoretische Erklärung. Denn da ein Akteur im Falle eines „prepaid incentive" den Anreiz bereits so oder so in Händen hält, wäre es für ihn nur rational, diese Vorleistung einzustreichen und die Gegenleistung schuldig zu bleiben. Dafür spricht insbesondere, dass der Akteur *nicht* davon ausgehen muss, dass eine Verweigerung der Kooperation in der Zukunft irgendwelche Konsequenzen für ihn hätte. Der Fall ist hier grundsätzlich anders gelagert als im Kontext sozialer Interaktionen, bei denen Ego davon auszugehen hat, in der Zukunft erneut auf Alter zu treffen bzw. treffen zu können. Erst dieses iterative Element schafft aber die Gelegenheiten, bei der Alter Ego eine Verweigerung der Kooperation heimzahlen könnte, Ego also zu befürchten hätte, dass ihm Alter die Verweigerung einer Kooperation künftig heimzahlen könnte.

5 „V-known model"

6 Fragebogen-/Interviewlänge: „kurz" bei Fragebögen mit bis zu 4 Seiten bzw. 31 Fragen bzw. bei bis zu 10 Minuten Interviewdauer; „mittel" bei Fragebögen mit 5 - 12 Seiten oder bis zu 75 Fragen bzw. bei bis zu 45 Minuten Interviewdauer; „lang" bei Fragebögen mit mehr als 12 Seiten bzw. bei mehr als 45 Minuten Interviewdauer. Vgl. Schnabel, Christiane (2005) Grundlagen der Survey Partizipation. Recherche und Metaanalyse. Bremen

Macht man sich nun aber eine auf das Gegenseitigkeitsprinzip abhebende austauschtheoretische Erklärung zu eigen, bleibt die Frage, worin im vorliegenden Fall die Wirkung dieses Prinzips begründet liegt: in seinem ökonomischen oder moralischen Gehalt? Liegt die Wirkung eines Anreizes also in seinem ökonomischen Wert oder in der symbolischen Geste als „kleiner Aufmerksamkeit" oder Dankeschön dafür, dass sich die Person der Bearbeitung und Rücksendung des Fragebogens annimmt? Folgen wir hier der Interpretation *Dillman's*[7], so würde der monetäre Anreiz gerade an Wirkung verlieren, wenn er im ökonomischen Sinne als Gegenleistung für die aufzuwendende Zeit bzw. Mühe wahrgenommen wird: Denn erhält der monetäre Anreiz die Konnotation, jemanden dafür zu bezahlen, dass er oder sie den Fragebogen bearbeitet, dann begünstigt dies die Haltung, den angebotenen Betrag danach zu beurteilen, ob er die erforderliche Mühe auch lohnt. Erscheint das Angebot dabei als inadäquate Gegenleistung für den zu betreibenden Aufwand, fällt es der Person leicht, das Angebot zurückzuweisen. Es ist dann wie bei einem Vertragsangebot, das zurückgewiesen wird, weil es der Akteur als für ihn nachteilig empfindet.

Nun erscheint die Frage ökonomischen vs. sozialen Austausches vor dem Hintergrund vorliegender Analyse keinesfalls so klar entscheidbar zu sein. Für eine „ökonomische" Interpretation des Gegenseitigkeitsprinzips würde sprechen, dass die Anreizstärke mit der Anreizhöhe steigt. Dieser Befund erwies sich aber unter den oben beschriebenen Bedingungen als nicht replizierbar. Für eine ökonomische Interpretation würde zudem der Befund sprechen, dass die Fragebogen- bzw. Interviewlänge den Anreizeffekt tendenziell kompensiert. Dass im Kontext der Incentivierungsfrage für beide Alternativen empirische Evidenz gefunden werden kann, zeigt sich auch bei Befunden aus zwei weiteren Studien.

Teilnahmebereitschaft in einem Access Panel

In einer dieser beiden Studien sind wir den Gründen für die Teilnahme an einer Befragung unter der Rahmenbedingung eines Access Panels nachgegangen. Befragt wurden Mitglieder des Online Access Panels „Sozioland" der Firma Globalpark (Köln)[8]. Die Befragung wurde im Juni 2005 als Online-Befragung realisiert. Zur Teilnahme eingeladen waren 5.430 Personen, von denen 40,1% (2.176 Personen) an der Befragung teilnahmen[9]. Diese fragten wir u.a., unter welchen Bedingungen es für die befragte Person eher unwahrscheinlich oder eher wahrschein-

7 Dillman, Don A. (2000) Mail and Internet Surveys. The Tailored Design Method. New York: Wiley, pp. 167 – 170.

8 Wir bedanken uns bei der Fa. Globalpark für die Möglichkeit, die Befragung durchführen zu können.

9 Pötschke, Manuela/Uwe Engel (2005) Datengüte in Online Access Panels: Determinanten der Teilnahmebereitschaft. Arbeitspapier. http://www.sozialforschung.uni-bremen.de/determinante.pdf

lich sei, an einer Befragung teilzunehmen, wenn sie dazu eingeladen werde (9-stufige Skala von „völlig unwahrscheinlich“ bis „sehr wahrscheinlich“). Eine Conjoint-Analyse zum Merkmal Incentivierung erbrachte die in Tabelle 3 ausgewiesenen Teilnutzenwerte:

Tabelle 3: Durchschnittl. Teilnutzenwerte zum Merkmal „Incentivierung“[10]

Anreizbedingung	
Weder Verlosung noch Belohnung	-0,39
Verlosung eines größeren Geldbetrages	-0,05
Verlosung wertvoller Sachpreise	-0,03
Ein kleines Dankeschön für jeden Eingeladenen	0
Ein kleines Dankeschön nach der Befragung	0,05
Geldbetrag für jeden vor der Befragung	0,09
Belohnung von 5 € für jeden Teilnehmer	0,11
Belohnung von mehr als 5 € für jeden Teilnehmer	0,21

Danach ist die Wirkung der Anreizbedingung auf die antizipierte Teilnahme negativ, wenn kein Anreiz geboten wird; sie ist tendenziell auch bei den beiden Verlosungsbedingungen negativ. Ein kleines Dankeschön erzielt nur eine positive Abweichung von der Nulllinie, wenn es nach der Befragung und somit nicht „prepaid“ gewährt wird. Positiv wirkt „Geld“ als Anreiz, und zwar in Abhängigkeit von der Höhe des Geldbetrages: Mehr als 5 € erzielen eine stärkere Wirkung als 5 €.

Offenbar wird also eine Gegenleistung erwartet, offenbar sollte dies eher ein monetärer Anreiz sein und offenbar ist mit einem höheren Anreizbetrag ein stärkerer Effekt zu erzielen als mit einem geringeren Betrag. Dieser Befund unterstützt im vorliegenden Zusammenhang sicherlich zwei Aussagen: Erstens, dass es sinnvoll ist, die Wirkung einer Anreizbedingung über das Gegenseitigkeitsprinzip zu erklären, und zweitens, dass dieses Prinzip eher in seiner ökonomischen als sozialen Variante zu wirken scheint.

Allerdings dürfen dabei die Grenzen der Aussagefähigkeit dieses Befundes nicht übersehen werden: es geht um erwartete Incentivierung bzw. darum, wovon die Wahrscheinlichkeit einer zukünftigen Teilnahme abhängt, also um Wirkungen auf eine *antizipierte* Teilnahme. Zudem kann sich der Befund nur auf diejenigen stützen, die sich an vorliegender Befragung selbst beteiligt haben. Das grenzt die Aussagefähigkeit in zweifacher Hinsicht ein: so bleibt unklar, ob sich der Befund replizieren lässt bzw. in welcher Weise er sich verändern würde, wenn a)

10 Quelle: Pötschke/Engel (2005, op.cit.), S. 8 (Diagramm 4)

auch die Nichtrespondenten zu vorliegender Befragung in die Analyse einbezogen werden könnten, und wenn b) außerhalb eines Access Panels und somit nicht nur in einem Bevölkerungsteil gefragt werden würde, der mit der Zustimmung zur Teilnahme *am Access Panel selbst* schon als grundsätzlich befragungsbereiter Bevölkerungsteil anzusehen ist.

So entsteht beispielsweise auch ein etwas anderes Bild, wenn wir uns auf Daten einer Studierendenbefragung stützen, die wir im November 2004 im Rahmen unseres „Studienbarometerprojekts“[11] an der Universität Bremen durchgeführt haben. Dabei ging es u.a. um Determinanten der Bereitschaft, einem Online Access Panel beizutreten, und darin eingebettet auch um die Frage, inwieweit bzw. in welcher Form eine Gegenleistung für die Beantwortung eines Fragebogens erwartet wird. Für den Fall, *dass* eine Gegenleistung erwartet wird, sollte die angemessenste Form benannt werden. Vorgegeben waren dazu die hauptsächlich eingesetzten Anreizformen Teilnahme an einer Verlosung, kleines Präsent und Geld.

Es zeigte sich, dass über zwei Drittel der 209 Befragten eher keine Gegenleistung erwarteten und dass weitere 13% dies nicht anzugeben wussten. 3½ Prozent fänden die Teilnahme an einer Lotterie am angemessensten, ca. 5 Prozent ein kleines Präsent und knapp 7½ Prozent Geld. Dabei teilten sich diejenigen, die sich für eine monetäre Gegenleistung aussprachen, in zwei Teilgrupen, und zwar je nachdem, ob sie eher für den Zeitaufwand honoriert werden wollten (ökonomische Variante: 4%) oder in dem Geld eher eine symbolische Geste, also die sprichwörtlich kleine Aufmerksamkeit, sahen (soziale Variante: 3½ %). Vergleichen wir diese beide Varianten in Bezug auf den Geldbetrag, der den Befragten für das Ausfüllen am angemessensten erschiene, so zeigte sich, dass in der „sozialen Variante“ die Wahl mehrheitlich auf 5 € fiel, während sich der Schwerpunkt in der „ökonomischen Variante“ auf einen höheren Geldbetrag verschob.

Der hohe Anteil von Befragten, die keine Gegenleistung erwarteten, scheint auf den ersten Blick die Angemessenheit einer austauschtheoretischen Erklärung in Frage zu stellen. Allerdings zeigte eine logistische Regression, die wir im Rahmen derselben Studie zu möglichen Determinanten eines *Adresspoolbeitritts* durchführten, unter anderem, dass die Auffassung, nicht „…ohne entsprechende Gegenleistung wertvolle Informationen zu liefern“ die Wahrscheinlichkeit verstärkte, seine E-Mail-Adresse für den Adresspool des Studienbarometers zur Verfügung zu stellen. Solche Gegenleistungen können natürlich auch *immaterieller* Art sein. Jedenfalls scheint die Bereitschaft zur Adresshergabe auch im Vertrauen darauf zu entstehen, dass es *in der einen oder anderen Form* eine Gegenleistung geben wird[12].

---

11 Quelle: Engel, Uwe/Manuela Pötschke/Sabine Scholz (2005) Abschlussbericht zum Projekt „Studienbarometer“.
http://www.sozialforschung.uni-bremen.de/studienbarometer_bericht.pdf, S. 31f.

12 Vgl. Engel/Pötschke/Scholz (2005, op.cit.), S. 28-34.

Natürlich können die hier vorgestellten Ergebnisse, die wir im Kontext des „Sozioland“ Access Panels und des Studienbarometerprojekts ermittelten, nur erste Einzelbefunde darstellen. Zu beachten ist auch, dass ihr Analysegegenstand keine objektiven Teilnahmewahrscheinlichkeiten darstellte, wie sie auf der Basis einer auch Nichtrespondenten einbeziehenden Analyse als Funktion möglicher Determinanten geschätzt werden könnten. Das Studienbarometerprojekt wendete sich im Übrigen an eine spezielle Population. Die beiden Studien können daher auch nur mögliche Anhaltspunkte zur Frage liefern, warum Anreize wirken. Weitere Studien zur Thematik erscheinen dringend erforderlich.

Zur Datenqualität der Bildungsangaben im Mikrozensus am Beispiel des Besuchs der gymnasialen Oberstufe und des allgemeinen Schulabschlusses

Bernhard Schimpl-Neimanns

Zusammenfassung

In diesem Beitrag wird am Beispiel des Besuchs der gymnasialen Oberstufe die Qualität der Bildungsangaben im Mikrozensus 1996 diskutiert. Ergänzend zum Schulbesuch werden erstmals auf Basis des Mikrozensuspanels Analysen zur Antwortkonsistenz der Angaben zum allgemeinen Schulabschluss vorgestellt. Die Vergleiche mit der amtlichen Bildungsstatistik zum Schulbesuch weisen auf Erhebungs- und Abgrenzungsprobleme im Mikrozensus hin. Erstens wird deutlich, dass Schüler allgemein bildender Schulen und beruflicher Schulen nicht den Definitionen des Mikrozensus entsprechend unterscheidbar sind. Zweitens ist eine gravierende Übererfassung bei den unter 18-jährigen Oberstufenschülern festzustellen, die auf eine problematische Unterscheidung der Klassenstufen 5-10 vs. 11-13 bzw. der Sekundarstufen I und II verweist. Die mit dem Mikrozensuspanel berechneten Übergangsraten der Bildungsabschlüsse zwischen verschiedenen Zeitpunkten weisen i. d. R. eine Stabilität von über 80 Prozent auf. Im Vergleich zu Ergebnissen sozialwissenschaftlicher Umfragen spricht dies für eine ausreichende bis gute Datenqualität des allgemeinen Schulabschlusses.

1 Einleitung[1]

Für die empirische Sozial- und Wirtschaftsforschung zählt der Mikrozensus zu den wichtigsten amtlichen Datenquellen für bildungsstatistische Analysen. Im Vergleich zur amtlichen Bildungsstatistik, die sich in die Schul-, Berufsschul- und Hochschulstatistik gliedert, lassen sich eine Reihe von Fragestellungen nur mit dem Mikrozensus bearbeiten. Dies betrifft insbesondere alle Sachverhalte, bei denen es um den Haushalts- und Familienkontext geht. Während beispielsweise in der Schulstatistik Angaben zur sozioökonomischen Lage der Eltern völ-

1 Für wertvolle Hinweise zu dem hier behandelten Thema danke ich Thomas Riede. Julia Schroedter danke ich für konstruktive Anmerkungen zu einer früheren Fassung.

lig fehlen, können diese Informationen für die noch bei ihren Eltern lebenden Schüler im Mikrozensus verwendet werden. Des Weiteren können mit dem Mikrozensus Aspekte der Verwertung von Bildungsqualifikationen auf dem Arbeitsmarkt oder Zusammenhänge zwischen Bildungsstatus und Lebenslage untersucht werden.

Für die Beurteilung von Analyseergebnissen des Mikrozensus sind, wie bei allen anderen Bevölkerungsumfragen, Informationen über die Datenqualität unabdingbar. Während zu den erwerbsstatistischen Themen des Mikrozensus solche Qualitätsuntersuchungen ansatzweise vorliegen (Dräther et al. 2001; Pöschl 1992; Rudolph 1998; Schupp et al. 1999), sind zu den bildungsstatistischen Angaben des Mikrozensus bislang kaum Einschätzungen möglich. Zu den wenigen Ausnahmen zählen insbesondere Untersuchungen zu den Auswirkungen der Freiwilligkeit der Auskünfte auf die Qualität der Bildungsangaben (Esser et al. 1989; Riede/Emmerling 1994), in Bezug auf den Hochschulabschluss frühere Vergleiche des Mikrozensus mit der Hochschulstatistik (Esser et al. 1989: 131) sowie zur beruflichen Weiterbildung ein Vergleich mit dem Berichtssystem Weiterbildung (Lois 2005).

Verteilungen des Mikrozensus dienen aufgrund des Stichprobenumfangs und der Auskunftspflicht für viele andere Stichproben in und außerhalb der amtlichen Statistik als Hochrechnungsrahmen bzw. werden für Prüfungen selektiver Befragungsausfälle, die häufig eng mit der Bildungsqualifikation der ausgewählten und zu befragenden Personen verbunden sind, herangezogen. Vor diesem Hintergrund der Verwendung des Mikrozensus als Referenzstatistik stellen sich deshalb besondere Anforderungen an die Datenqualität.

Für die Qualitätsbeurteilung liefern Panelangaben wichtige Informationen. Beispielsweise kann auf die Angaben aus vorherigen Befragungen zurückgegriffen werden. Damit werden unter anderem Fragen zur Antwortkonsistenz bearbeitbar. Seit dem Mikrozensusgesetz 1996 dürfen die statistischen Ämter die für die Zusammenführung der Querschnittsangaben zu einem Rotationspanel benötigten Ordnungsnummern speichern. In einem vom Bundesministerium für Bildung und Forschung (BMBF) und der Deutschen Forschungsgemeinschaft (DFG) finanzierten Projekt werden die Voraussetzungen für die Weitergabe des Mikrozensus als Rotationspanel an die Wissenschaft geschaffen.[2] Für die Einschätzung des Analysepotenzials wurden im Projekt unter anderem Bildungsverläufe untersucht. Aus diesem Projekt werden einige Befunde zur Datenqualität der Angaben zum Besuch der gymnasialen Oberstufe aufgegriffen (Schimpl-Neimanns 2005). Ergänzend zum Schulbesuch wird die zeitliche Konsistenz der Angaben zum allgemeinen Schulabschluss berichtet.

In den bisherigen Untersuchungen zum Besuch der gymnasialen Oberstufe wurde deutlich, dass Schüler bzw. Absolventen allgemein bildender Schulen und

2 Siehe dazu die WWW-Seite des Projekts www.destatis.de/mv/mzpanel_start.htm.

beruflicher Schulen nicht den Definitionen entsprechend unterscheidbar sind. Um den vermuteten Abweichungsgründen nachzugehen, werden in diesem Aufsatz zusätzliche Vergleiche mit der Bildungsstatistik vorgenommen. Seit 1991 wird der Schulbesuch in Anlehnung an die Internationale Standardklassifikation für das Bildungswesen (ISCED) im Mikrozensus nach Klassenstufen erhoben. Da der Schulbesuch bis 1990 nach Schularten erfasst wurde, kann insbesondere der Vergleich von Ergebnissen des Mikrozensus 1989 mit der Bildungsstatistik darüber Aufschluss geben, ob und in welcher Weise die bei den Mikrozensen 1996 bis 1999 gefundenen Fehlklassifikationen auf im Alltag eher schwierige Begriffe zurückzuführen sind.

Im Folgenden werden zunächst die verwendeten Daten und die Methode beschrieben sowie Gründe für die bei dem Vergleich der Datenquellen auftretenden möglichen Abweichungen genannt. Im vierten Abschnitt werden Verteilungen zum Besuch der gymnasialen Oberstufe im Mikrozensus 1996, 1991 und 1989 den Ergebnissen der Bildungsstatistik gegenüber gestellt. Anschließend werden erste Ergebnisse des Mikrozensuspanels 1996-1999 zur Antwortstabilität des allgemeinen Schulabschlusses berichtet. Der Aufsatz schließt mit einer zusammenfassenden Einschätzung zu den Verteilungsabweichungen.

2 Datenbeschreibung

Seit 2005 wird der Mikrozensus als unterjährige Erhebung durchgeführt. Mit dieser Umstellung auf eine kontinuierliche Befragung sind weitere Modifikationen des Frageprogramms sowie methodische Änderungen verbunden (siehe Afentakis/Bihler 2005; Lotze/Breiholz 2002a,b). Die folgende Darstellung bezieht sich auf die verwendeten Daten des Mikrozensus 1989 bis 1996 mit Konzentration auf das von 1990 bis 2004 geltende Stichprobendesign und Frageprogramm. Im Anschluss an die Kurzbeschreibung der Bildungsstatistik wird die Konstruktion vergleichbarer Bildungsdaten skizziert.

2.1 Mikrozensus

Im Mikrozensus werden seit 1957 in Westdeutschland und seit 1991 in den neuen Bundesländern mit einem Stichprobenumfang von einem Prozent der Personen und Haushalte jährlich vielfältige Informationen über die demografische, soziale und wirtschaftliche Struktur der Bevölkerung erhoben. Aufgrund der Auskunftspflicht liegt die Teilnahmequote der Haushalte bei rund 97 Prozent. In Bezug auf das Stichprobendesign ist der Mikrozensus als mehrfach geschichtete einstufige Klumpen- bzw. Flächenstichprobe und als Rotationspanel gekennzeichnet (Meyer 1994). Die Primäreinheiten bilden über 40.000 Auswahlbezirke (Klumpen). Sie bestehen in den Mikrozensen ab 1990 aus benachbarten Wohnungen, die auf

Basis von Ergebnissen der Volkszählung 1987 bzw. des Zentralen Einwohnerregisters in den neuen Bundesländern gebildet wurden. Die Auswahlbezirke umfassen durchschnittlich etwa neun Wohnungen; im Mikrozensus bis 1989 waren es noch rund 23 Wohnungen. Die jährliche Aktualisierung der Mikrozensus-Stichprobe berücksichtigt die neu entstandenen Wohnungen bzw. Bebauung ganzer Flächen durch Verwendung von Daten der Bautätigkeitsstatistik.

Der Mikrozensus ist als eine Panelstichprobe angelegt, bei der die Haushalte eines Auswahlbezirkes vier Jahre lang befragt werden, wobei jedes Jahr ein Viertel der Auswahlbezirke ausgetauscht wird. Wegziehende Personen und Haushalte werden jedoch nicht weiter befragt, sondern durch die nachziehenden Personen bzw. Haushalte ersetzt (Prinzip der Flächenstichprobe).

In Bezug auf die Befragungsmethode dominiert die persönliche Befragung. Rund 15 Prozent der Interviews basieren auf schriftlichen Auskünften der Befragten. Fremdauskünfte (Proxy-Interviews) sind möglich, z. B. wenn Eltern die Fragen zum Schulbesuch und Bildungsabschluss ihrer Kinder beantworten. Erst ab der Erhebung 1999 wird im Rahmen der Unterstichprobe des Mikrozensus erfasst, ob es sich um Selbst- oder Fremdauskünfte handelt. Insgesamt beruhen knapp 30 Prozent der Angaben auf Proxy-Interviews (Breiholz 2000). Bis zum Mikrozensus 2004 bezieht sich der Berichtszeitraum i. d. R. auf die letzte feiertagsfreie Woche im April eines Jahres (Berichtswoche).

Die meisten Fragen sind auskunftspflichtig. Daneben gibt es weitere Fragen und Themenbereiche, die von der Auskunftspflicht ausgenommen sind. Hierzu zählen beispielsweise Fragen der EU-Arbeitskräftestichprobe (European Labour Force Survey), die als Unterstichprobe in den Mikrozensus integriert ist, Fragen zur Gesundheit sowie (seit 1996) die Angaben zum allgemeinen und beruflichen Abschluss von Personen über 50 Jahren.

Die Fallzahlen des Mikrozensus werden in einem zweistufigen Verfahren hochgerechnet. Auf der ersten Stufe erfolgt eine Korrektur des Unit-Nonresponse von rund drei Prozent für die nicht erreichten Haushalte. Auf der zweiten Stufe werden die nach dieser Ausfallkorrektur gewichteten Mikrozensusergebnisse an bekannte demografische Randverteilungen angepasst. Diese Anpassung an Ergebnisse der laufenden Bevölkerungsfortschreibung erfolgt für die Merkmale Geschlecht und Staatsangehörigkeit (Deutsche vs. Ausländer) auf der regionalen Ebene von so genannten Anpassungsschichten sowie für Soldaten und Wehrpflichtige an entsprechende Bestandsmeldungen (siehe Heidenreich 1994). Die Anpassungsschichten umfassen durchschnittlich wenigstens 500.000 Einwohner.

Für die folgenden Auswertungen werden Daten der Mikrozensus Scientific Use Files verwendet. Dabei handelt es sich jeweils um faktisch anonymisierte 70-Prozent-Haushaltssubstichproben des vollen Mikrozensus. Das Ziehungsverfahren berücksichtigt die wesentlichen Designelemente des Mikrozensus und ge-

währleistet eine sehr hohe Übereinstimmung mit Ergebnissen der Originaldaten (Rendtel/Schimpl-Neimanns 2001). Das verwendete Panelfile des Mikrozensus umfasst die im Zeitraum von 1996 bis 1999 befragten Personen (Rotationsviertel). Die hier genutzten faktisch anonymisierten Paneldaten entsprechen zirka einer 65-Prozent-Haushaltssubstichprobe.

2.2 Bildungsstatistik

Die Statistik zu allgemein bildenden Schulen basiert auf einer Vollerhebung und ist als Teil der amtlichen Bildungsstatistik in den einzelnen Bundesländern unterschiedlich organisiert. Durch die Orientierung an einem Minimalprogramm können aber Bundesergebnisse zusammengestellt werden (Statistisches Bundesamt 1996; Weishaupt/Fickermann 2001). Die Statistik beruht auf Meldungen der Schulleiter (Anstaltsbefragung). Räumlichen Darstellungen, wie beispielsweise dem Schulbesuch nach Bundesland, liegt der Schulort zugrunde. Die summarischen Angaben umfassen unter anderem die Zahl der Schüler nach Klassenstufe, Alter, Geschlecht und Staatsangehörigkeit. Dass Klassenstufe und Alter nicht in Kombination ausgewiesen werden, schränkt Vergleiche mit dem Mikrozensus ein. Zu beachten ist auch, dass zur Verringerung der Belastung der Auskunft gebenden Schulen der Schulbesuch nicht in jedem Jahr und nicht in jedem Bundesland differenziert erhoben wird. Beispielsweise werden die nach Geburtsjahr gegliederten Schülerzahlen nur im Abstand mehrerer Jahre ermittelt. Die Altersverteilungen der dazwischen liegenden Zeitpunkte werden auf Basis der ausführlichen Strukturerhebungen geschätzt.

Bei der Schulstatistik handelt es sich um Stichtagserhebungen, deren Berichtszeitraum sich jeweils auf den Beginn eines Schuljahres (Herbst) bezieht. Die Angaben über Schulentlassene nach Abschlussart beziehen sich auf das vergangene Schuljahr bzw. das Schuljahresende.

2.3 Vergleichbare Erfassung der gymnasialen Oberstufe

In den alten Bundesländern und Brandenburg bilden bis Ende der 1990-er Jahre die Klassenstufen 11-13 allgemein bildender Schulen der Sekundarstufe II die gymnasiale Oberstufe. In Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen kann das Abitur schon nach 12 Schuljahren absolviert werden. Der klassische Abschluss der gymnasialen Oberstufe ist die allgemeine Hochschulreife (Abitur) oder die fachgebundene Hochschulreife. In einzelnen Bundesländern wird bei vorliegendem Versetzungszeugnis in die 13. (bzw. 12.) Klasse der Abschluss der Fachhochschulreife vergeben, die eigentlich ein beruflicher Bildungsabschluss ist. Neben der Fachhochschulreife können sowohl die fachgebundene Hochschulreife als auch die allgemeine Hochschulreife auch an beruflichen Schulen erlangt werden (KMK 2003). In den meisten Bundesländern zählen

zur gymnasialen Oberstufe Gymnasien, Integrierte Gesamtschulen und Freie Waldorfschulen, daneben die Einrichtungen Kolleg und Abendgymnasium des zweiten Bildungsweges. Die gymnasiale Oberstufe ist in den einzelnen Bundesländern unterschiedlich organisiert (Bellenberg et al. 2004). Während die Kultusministerkonferenz in ihrer Übersicht des deutschen Schulsystems bei der Abgrenzung der gymnasialen Oberstufe eine breite Definition verwendet (KMK 2001), die auch berufliche Gymnasien einschließt, wird sowohl im Mikrozensus als auch in der Schulstatistik strikt zwischen allgemein bildenden und beruflichen Schulen unterschieden. Berufliche Gymnasien mit dem Bildungsziel Abitur oder Fachhochschulreife (Fachoberschulen, Fachgymnasien, Berufs-/Technische Oberschulen) gehören demnach zu den beruflichen Schulen.

Seit 1991 wird im Mikrozensus die Filterfrage nach dem gegenwärtigen Schulbesuch gestellt. Schulbesucher werden nach der Art der Schule gefragt, wobei Schüler einer allgemein bildenden Schule die Klassenstufe (1-4, 5-10, 11-13) angeben können. Die Kategorie „Klassenstufe 11-13“ ist in Klammern mit dem Zusatz „gymnasiale Oberstufe“ versehen. An dieser Stelle ist anzumerken, dass Befragte, die sich für eine schriftliche Beantwortung mittels Selbstausfüller-Fragebogen entschieden haben, im Fragebogen keine weiteren Erläuterungen zu den abgefragten Begriffen erhalten.[3] Eine Gegenüberstellung der Schularten zur gymnasialen Oberstufe bzw. zu den beruflichen Schulen findet sich nur im Interviewerhandbuch. Da rund 85 Prozent aller Angaben auf klassischen Interviews beruhen, kommt es daher in erster Linie darauf an, ob der Interviewer die genauen Definitionen kennt und anwendet.

Bis 1990 wurde der Schulbesuch im Mikrozensus ohne die Unterscheidung allgemein bildender und anderer Schulen und nicht nach Klassenstufen erhoben, sondern es wurde die jeweilige Schulform erfragt. Dabei waren Schüler aller Einrichtungen mit dem Bildungsziel Fachhochschulreife und Hochschulreife in der Kategorie „Gymnasium/Fachoberschule“ anzugeben. Im Unterschied zu den Erhebungen ab 1991 sind damit 1989 berufliche Gymnasien den allgemein bildenden Schulen der Sekundarstufe II zugeordnet. Der Besuch der Gymnasialzüge an Gesamt- und Sonderschulen (Sekundarstufe II) war ebenfalls in der Kategorie „Gymnasium/Fachoberschule“ einzutragen. Die genauen Zuordnungen werden wiederum lediglich im Interviewerhandbuch beschrieben.

Die im Mikrozensus verwendeten Definitionen stimmen weitestgehend mit der Schulstatistik überein bzw. die jeweiligen Schulformen können vergleichbar abgegrenzt werden. Für den Vergleich des Mikrozensus mit der Schulstatistik ist dies wichtig, da somit konzeptionelle Abweichungen ausgeschlossen werden können.

3 Die Fragebögen sind unter www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/index.htm zugänglich.

3 Methode und Abweichungsgründe

Datenqualität bezieht sich allgemein darauf, in welcher Weise statistische Ergebnisse den mit der Datenerhebung verfolgten Zweck erfüllen. Zu den wichtigsten Aspekten zählen Schätzgenauigkeit und Reliabilität sowie Validität, aber auch die Verfügbarkeit und der Datenzugang (Office of Management and Budget 2001: 1-2). Häufig steht bei Fragen der Datenqualität die Genauigkeit bzw. der Stichprobenfehler im Vordergrund. Jedoch dürfen bei der Frage der Datenqualität systematische Fehler keinesfalls ausgeblendet werden, da sie in vielen Umfragen größer als der Stichprobenfehler sein können (Särndal et al. 1997: 539).

Die Aufdeckung systematischer Fehler ist nicht einfach und setzt oftmals externe Datenquellen voraus. Im Idealfall werden verschiedene Verfahren kombiniert. Dazu gehören Pretests zur Kontrolle des Fragebogens und Wiederholungsbefragungen mit anderen Frageversionen bzw. differenzierten Nachfragen zu einzelnen Sachverhalten, die zugleich der Interviewerkontrolle dienen können. In wenigen Fällen sind Verknüpfungen von Umfragedaten mit Registerdaten und ein Abgleich der in beiden Datenquellen erfassten Merkmale möglich. Häufiger werden Randverteilungen mit anderen Daten verglichen. Bei den standardgemäß durchgeführten Analysen US-amerikanischer statistischer Ämter zur Datenqualität kommen diese Methoden häufig zum Einsatz (siehe unter anderem: Black et al. 2003; Brick et al. 1996; McGovern/Bushery 1999; NCES 1997; U.S. Census Bureau 2004).

Im Folgenden werden lediglich Randverteilungsvergleiche, also einfache deskriptive Vergleiche hochgerechneter Mikrozensusergebnisse mit der Bildungsstatistik durchgeführt. Dabei wird nicht vorausgesetzt, dass die Bildungsstatistik „wahre“ Werte liefert. Allerdings wird angenommen, dass die Ergebnisse der Totalerhebung Bildungsstatistik weniger fehlerbehaftet sind als die des Mikrozensus. Diese Annahme erscheint gerechtfertigt.

Als Quellen systematischer Fehler kommen unklare Fragen oder Filterführungen im Fragebogen, Probleme der Erhebungs- bzw. Befragungsmethode, Interviewereinflüsse oder Verständnisschwierigkeiten und fehlerhafte Auskünfte sowie Antwortausfälle vonseiten der Befragten in Frage. Aus dem Vergleich mit der Bildungsstatistik kann nur die Gesamtabweichung („Total Survey Error“; siehe Biemer/Lyberg 2003) ermittelt werden. Inwiefern die Gesamtabweichungen mit dem Stichprobenfehler, unterschiedlichen Berichtszeiträumen, Unit- und Item-Nonresponse, verschiedenen Antwortfehlern (u. a. aufgrund sozialer Erwünschtheit) oder unterschiedlichen Befragungsarten (Mode-Effekte) zusammenhängen, kann nicht geklärt werden. Zur ansatzweisen Kontrolle werden, sofern möglich, weitere Differenzierungen vorgenommen.

Die den Gesamtabweichungen zwischen Mikrozensus und Bildungsstatistik zugrunde liegenden Ursachen können wie folgt zusammengefasst werden:

- **Berichtszeitraum:** Während die Bildungsstatistik die Schülerzahlen zum Schulanfang eines Schuljahres (Herbst) berichtet, beziehen sich die Mikrozensusergebnisse auf die Berichtswoche des Folgejahres, die i. d. R. in der letzten feiertagsfreien Aprilwoche liegt. Die Unterschiede, die auf die Differenz zwischen Herbst des Vorjahres und April zurückzuführen sind, dürften jedoch gering sein, da Schulartwechsel und Ausbildungsabbrüche hauptsächlich zum Schuljahresende stattfinden.
- Zu berücksichtigen ist aber, dass im Mikrozensus nach dem „gegenwärtigen" Schulbesuch gefragt wird. Bei einem bis in den Sommer reichenden Befragungszeitraum kann die Fokussierung auf den Interviewzeitpunkt dazu führen, dass Schüler, die in den Schulferien, d. h. nach Abschluss des Schuljahres befragt werden, Angaben zum nächsten Schuljahr (Herbst) statt zum vergangenen Schuljahr machen. Gegebenenfalls wird auch der zwischenzeitlich erreichte Schulabschluss statt der noch im April besuchten Schule angegeben. Grundsätzlich nicht auszuschließen ist auch der Fall, dass Angaben exakt zum Befragungszeitpunkt erfolgen, d. h. eventuell kein Schulbesuch berichtet wird. Allerdings ist damit kaum zu rechnen, sofern von den befragten Haushalten auch in den Schulferien der Hauptstatus eines Schülers wahrgenommen wird.
- **Merkmalsunterschiede bei Drittvariablen:** Drittvariablen liegen in der Bildungsstatistik nur sehr begrenzt vor. In dem Zusammenhang, dass in der Bildungsstatistik das Alter bzw. das Geburtsjahr nicht jedes Jahr direkt erhoben, sondern entsprechende Altersverteilungen teilweise geschätzt werden (s. o.), kann es zu Abweichungen zum Mikrozensus kommen. Diese dürften jedoch eher marginal sein.
- **Populationsschätzung, Hochrechnung:** Frühere Vergleiche zwischen Mikrozensus und Volkszählung haben gezeigt, dass die laufende Bevölkerungsfortschreibung, an die die Mikrozensusdaten angepasst werden, den Bevölkerungsbestand mit zunehmendem Abstand zur Volkszählung überschätzt (siehe zusammenfassend Rendtel/Schimpl-Neimanns 2001; Statistische Ämter des Bundes und der Länder 2004). Durch die Anpassung können somit Fehler in den Mikrozensus übertragen werden. Die Abweichungen zu den nicht extra angepassten Ergebnissen der Bildungsstatistik sollten sich bei Kontrolle weiterer Variablen, die wie beispielsweise das Alter nicht in der Anpassung berücksichtigt werden, jedoch in Form einer konstanten Verschiebung zeigen.
- **Definitionen im Mikrozensus:** Wie oben angesprochen, zählen im Mikrozensus ab 1991 ausschließlich Schüler allgemein bildender Schulen der Klassenstufen 11-13 zur gymnasialen Oberstufe. Schüler beruflicher Schulen mit dem Bildungsziel Fachhochschulreife oder allgemeine Hochschulreife sind dagegen als Schüler beruflicher Schulen einzuordnen. Es ist davon auszugehen,

dass diese Definition nicht dem Alltagsverständnis entspricht und deshalb in der Befragungspraxis nicht adäquat umgesetzt werden kann.

- **Befragungsart (Mode-Effekte):** Bei den Besuchern der gymnasialen Oberstufe liegt der Anteil der schriftlichen Auskünfte bei 14 Prozent. Da die genauen Definitionen nur im Interviewerhandbuch zu finden sind, nicht aber im Selbstausfüller-Fragebogen, ist damit zu rechnen, dass schriftlich Befragte bei der Beantwortung der Frage nach dem Schulbesuch auf ihre Alltagsdefinition zurückgreifen. Geht man aber davon aus, dass die Definitionen selbst im mündlichen Interview nicht immer umfassend berücksichtigt werden, dürften die unterschiedlichen Befragungsarten nur geringe Effekte besitzen.
- Bei Oberstufenschülern liegen im Mikrozensus 1999 mit 82 Prozent überdurchschnittlich häufig Fremdauskünfte vor. Hinsichtlich der generellen Einschätzung einer bei Proxy-Interviews eingeschränkten Datenqualität wäre anzunehmen, dass dies auch in Bezug auf die Angaben zum Schulbesuch zutrifft. Anders als bei Proxy-Angaben zum Einkommen oder zu Arbeitsstunden (Dawe/Knight 1997) kann man jedoch davon ausgehen, dass Eltern über die von ihren Kindern besuchte Schule gut informiert sind.

4 Vergleiche zum Besuch der gymnasialen Oberstufe im Mikrozensus und in der Bildungsstatistik

Wie eingangs erwähnt, wird als Hauptgrund für die Abweichungen zwischen Mikrozensus und Bildungsstatistik in Bezug auf Schüler der gymnasialen Oberstufe angenommen, dass diese auf Begriffe und Definitionen zurückzuführen sind, die sich in der Befragungspraxis nicht unmittelbar umsetzen lassen. Zur Überprüfung dieser Vermutung stehen daher Vergleiche des Mikrozensus 1996 mit der Bildungsstatistik im Zentrum dieses Abschnitts. Sie werden durch Vergleiche früherer Mikrozensusdaten ergänzt.

Gemäß der Bildungsstatistik besuchten im Schuljahr 1995/96 679.900 Schüler im Alter bis zu 21 Jahren die gymnasiale Oberstufe (siehe Abb. 1). Der Mikrozensus 1996 weist dagegen hochgerechnet 1.093.600 Schüler aus. Die Differenz von über 413.000 Schülern entspricht einer Übererfassung von insgesamt rund 60 Prozent. Bei den Gesamtwerten des Mikrozensus sind die 95 %-Konfidenzintervalle ausgewiesen.[4] Die Differenzen zur Bildungsstatistik sind offensichtlich nicht durch den Stichprobenfehler zu erklären. Die Überschätzung reduziert sich insgesamt auf ungefähr die Hälfte (32 %), wenn zu den Schülern der Sekundar-

4 Zur Berechnung der Stichprobenfehler bei der hier verwendeten gebundenen Hochrechnung (Anpassung an die Bevölkerungsfortschreibung) siehe Rendtel/Schimpl-Neimanns (2001: 103f.). Dieses Verfahren ist nur mit den Scientific Use Files des Mikrozensus ab 1996 anwendbar.

stufe II allgemein bildender Schulen der Bildungsstatistik die Schüler beruflicher Gymnasien hinzugerechnet werden. In diesem Fall schließt das Konfidenzintervall des Mikrozensus bei den 18- bis 21-Jährigen die entsprechenden Gesamtwerte der Bildungsstatistik ein. Dies deutet darauf hin, dass im Mikrozensus Besucher beruflicher Gymnasien entgegen den Definitionen des Mikrozensus als Schüler der gymnasialen Oberstufe erfasst sind. Für diese Annahme spricht auch, dass durch die Herausnahme der Schüler beruflicher Gymnasien (Fachoberschulen, Fachgymnasien und Berufs-/Technische Oberschulen) aus den Gesamtwerten der Berufsschulstatistik die Anpassung des Mikrozensus an die Berufsschulstatistik verbessert werden kann (hier ohne Nachweis; siehe Schimpl-Neimanns 2005: 28-29).

Bei den unter 18-Jährigen, insbesondere bei den unter 17-Jährigen bestehen dennoch weiterhin erhebliche Übererfassungen, die offensichtlich nicht mit der Fehlklassifikation beruflicher Gymnasien zu erklären sind. Diese Abweichungen könnten auf das oben genannte Problem des Berichtszeitraums bzw. auf die Fokussierung auf den Interviewzeitpunkt und Antworten zum Schulbesuchsstatus des neuen Schuljahres zurückzuführen sein. Sie könnten aber auch damit zusammenhängen, dass Befragten die Unterscheidung des Schulbesuchs nach Sekundarstufen (Klassenstufen 5-10 vs. 11-13) schwer fällt und die zur Klassenstufe 11-13 in Klammern genannte Ergänzung „gymnasiale Oberstufe“ allgemein als „Gymnasium“ missverstanden wird und von der Unterscheidung der Sekundarstufen ablenkt.[5] Da die Bildungsstatistik die Altersgruppen, bei denen Überschneidungen zwischen den Sekundarstufen I und II auftreten, nur zusammengefasst in nach oben bzw. unten offenen Flügelklassen ausweist, lassen sich diese Vermutungen nicht weiter verfolgen.

Die genannten Befunde zeigen sich auch für die Erhebungszeitpunkte 1997 bis 1999 (siehe Schimpl-Neimanns 2005: 56-58) und weisen somit auf ein systematisches Problem hin. Im dritten Abschnitt wurden als Ursachen auch andere potenzielle Zusammenhänge zwischen Befragungsart und Verteilungsabweichung angesprochen. Um diese versuchsweise zu kontrollieren, wurde eine logistische Regression zum Verhältnis des Besuchs der gymnasialen Oberstufe vs. einer beruflichen Schule mit soziodemografischen Variablen und den Befragungsarten Selbstausfüller und Proxy-Interview geschätzt. Mit den Daten des Mikrozensus 1999 ergaben sich keine statistisch signifikanten Interaktionen zwischen den Variablen der Befragungsart und den anderen erklärenden Variablen (hier ohne Darstellung). Damit kann zwar nicht ausgeschlossen werden, dass solche Effekte dennoch vorhanden sind. Für eine Überprüfung von Mode-Effekten liegen

5 Im Mikrozensus enthält die Flügelklasse der unter 17-Jährigen auch den Geburtsjahrgang 1981. Dass hochgerechnet rund 18.400 Schüler, die im Herbst 1995 etwa 14 bis 15 Jahre alt waren, einen Besuch der gymnasialen Oberstufe angeben, könnte als Indiz für eine unscharfe Differenzierung zwischen den Sekundarstufen I und II gewertet werden.

jedoch im Mikrozensus keine geeigneten Informationen vor. Von entscheidender Bedeutung wäre die Kenntnis des Interviewzeitpunkts.

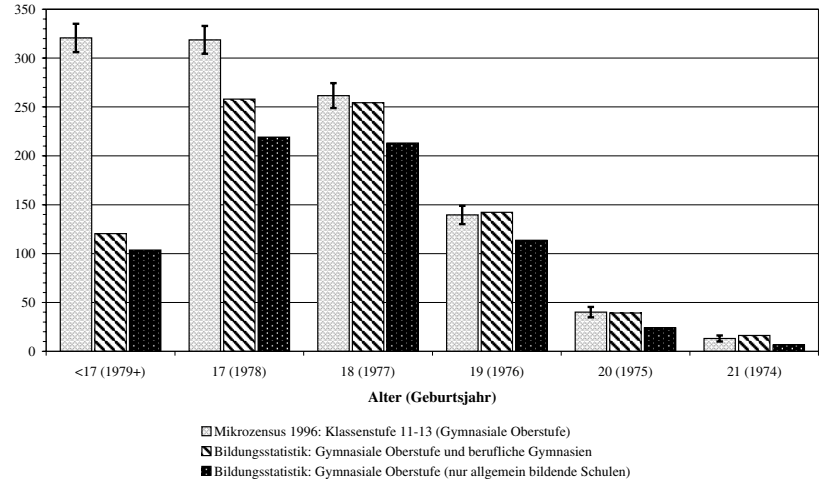


Abb. 1: Besucher der gymnasialen Oberstufe im Mikrozensus 1996 und in der Bildungsstatistik (Schuljahr 1995/96) in 1.000

Quellen: Bildungsstatistik: Gymnasiale Oberstufe allgemein bildender Schulen: Statistisches Bundesamt 1997a: 24, 36-37. Berufliche Schulen: Statistisches Bundesamt 1997b: 17. (Gymnasiale Oberstufe: N=679.900; gymnasiale Oberstufe incl. berufliche Gymnasien: N=830.300).
Mikrozensus 1996 (faktisch anonymisierte 70 %-Substichprobe), Bevölkerung am Ort der Hauptwohnung; hochgerechnete, an die Bevölkerungsfortschreibung angepasste Fallzahlen (N=1.093.600; ungewichtet: n=6.753); 95 %-Konfidenzintervall: |–|; eigene Berechnungen.

Um die Plausibilität der Erklärung der Verteilungsunterschiede trotz fehlender methodischer Informationen zu testen, kann auf den Mikrozensus 1989 zurückgegriffen werden. Bis 1990 wurde der Schulbesuch nach Schularten erfragt. Alle allgemein bildenden und beruflichen Schulen mit dem Bildungsziel allgemeine Hochschulreife (Abitur), fachgebundene Hochschulreife oder Fachhochschulreife wurden in der Kategorie „Gymnasium/Fachoberschule“ zusammengefasst. Trifft die Annahme zur Fehlklassifikation der beruflichen Gymnasien zu, ist zu erwarten, dass die Abweichungen zwischen Bildungsstatistik und Mikrozensus 1989 geringer ausfallen. Zum Vergleich werden zunächst die Mikrozensusergebnisse des Jahres 1991 herangezogen. Da im Mikrozensus 1989 die Abgrenzung der Sekundarstufe II nur näherungsweise mittels des Alters möglich ist, wird im Folgenden auf die nach unten offene Flügelklasse sowie auf die Altersgruppe der

21-Jährigen, die für einige Schulformen in der Bildungsstatistik ebenfalls zusammengefasste Jahrgänge enthält, verzichtet. Außerdem beziehen sich die Gegenüberstellungen auf Westdeutschland, d. h. nur auf das frühere Bundesgebiet.

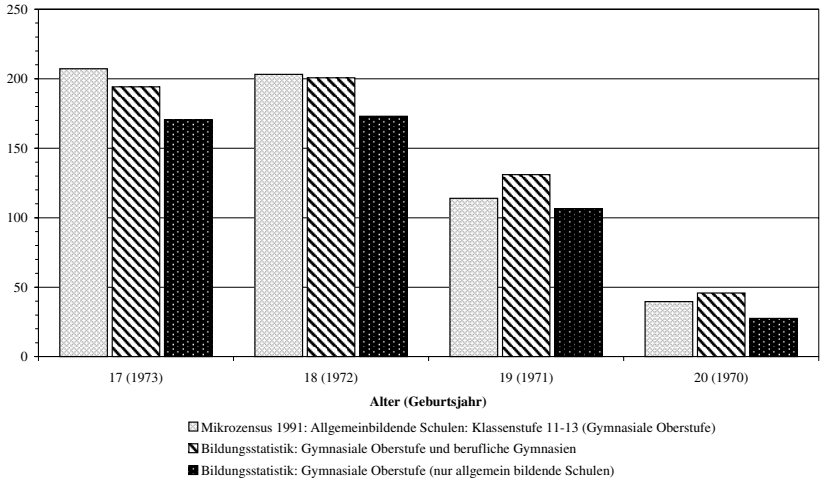


Abb. 2: Besucher der gymnasialen Oberstufe im Mikrozensus 1991 und in der Bildungsstatistik (Schuljahr 1990/91) in 1.000 - Westdeutschland

Quellen: Bildungsstatistik: Gymnasiale Oberstufe allgemein bildender Schulen (N=477.500); gymnasiale Oberstufe inkl. berufliche Gymnasien (N=571.500); Statistisches Bundesamt 1992.
Mikrozensus 1991 (faktisch anonymisierte 70 %-Substichprobe), Bevölkerung am Ort der Hauptwohnung; hochgerechnete, an die Bevölkerungsfortschreibung angepasste Fallzahlen (N=563.800; ungewichtet n=3.578); eigene Berechnungen.

Der Vergleich des Mikrozensus 1991 mit der Bildungsstatistik (siehe Abb. 2) zeigt in der Tendenz geringere, aber im Muster ähnliche Abweichungen wie beim Mikrozensus 1996.[6] Insgesamt wird die Zahl der gymnasialen Oberstufenschüler um 18 Prozent (= 563.800 / 477.500) überschätzt. Zur Gesamtabweichung trägt hauptsächlich die Übererfassung von rund 22 Prozent bei den 17-Jährigen bei.

6 Da die neuen Fragen zum Schulbesuch erstmals 1991 gestellt wurden, ist nicht auszuschließen, dass die Interviewer teilweise noch die alten Definitionen zum Schulbesuch angewendet haben. Dies könnte ein Grund für die 1991 im Vergleich zu 1996 geringeren Abweichungen zwischen Mikrozensus und Bildungsstatistik und das im Vergleich zum Mikrozensus 1989 ähnliche Abweichungsmuster sein.

Dies verweist wiederum auf die Probleme der Unterscheidung der Sekundarstufen bzw. die Fokussierung auf den Interviewzeitpunkt. Wie beim Mikrozensus 1996 sind, mit Ausnahme der 19-Jährigen, die altersspezifischen Abweichungen zur Bildungsstatistik geringer, wenn zu den Besuchern der gymnasialen Oberstufe auch die Schüler beruflicher Gymnasien gezählt werden. Bei den über 18-Jährigen kehrt sich allerdings die Übererfassung in eine Untererfassung von rund 13 Prozent um. Insgesamt ist die Differenz zwischen Mikrozensus und Bildungsstatistik von -1,3 Prozent vernachlässigbar, wenn man in der Bildungsstatistik die Schüler beruflicher Gymnasien hinzunimmt.

Mit dem Mikrozensus 1989 werden 547.700 Schüler im Alter von 17 bis 20 Jahren geschätzt, die Schulen mit dem Bildungsziel des Abiturs oder der Fachhochschulreife besuchen. Der hierzu vergleichbare Gesamtwert der Bildungsstatistik beträgt 609.400 Schüler. Er umfasst die der gymnasialen Oberstufe zugeordneten allgemein bildenden Schulen (siehe Abschnitt 2.3) sowie berufliche Gymnasien. Damit liegt im Mikrozensus 1989 insgesamt eine Untererfassung von 10 Prozent vor. Im Vergleich zur Übererfassung der gymnasialen Oberstufe im Mikrozensus 1991 (18 %) und insbesondere zur gravierenden Übererfassung dieser Gruppe im Mikrozensus 1996 (60 %) ist die (absolute) Abweichung des Mikrozensus 1989 vom Sollwert der Bildungsstatistik deutlich geringer. Dies unterstützt die obige Vermutung, dass die im Mikrozensus ab 1991 verwendete Abgrenzung der gymnasialen Oberstufe mit der Zuordnung beruflicher Gymnasien zu beruflichen Schulen in der Befragungspraxis schwer umzusetzen ist und eine wichtige Quelle der Verteilungsabweichungen zur Bildungsstatistik darstellt.

In Bezug auf die offene Frage, ob die Abweichungen zwischen dem Mikrozensus ab 1991 und der Bildungsstatistik bei den unter 18-Jährigen mit dem Problem der Fokussierung auf den Interviewzeitpunkt oder mit der schwierigen Unterscheidung der Sekundarstufen I und II (bzw. Klassenstufen 5-10 vs. 11-13) zusammenhängen, ist der Vergleich der Verteilungsunterschiede der Mikrozensen 1989 und 1991 zur Bildungsstatistik aufschlussreich. Verwendet man als Sollzahlen der Bildungsstatistik die Gesamtwerte der Schüler allgemein bildender Schulen der gymnasialen Oberstufe einschließlich beruflicher Gymnasien, zeigt sich für die 17-Jährigen folgender Befund: Während im Mikrozensus 1991 eine Übererfassung des Gesamtwerts der Bildungsstatistik um rund 7 Prozent vorliegt,[7] wird der Sollwert der Bildungsstatistik mit dem Mikrozensus 1989 nur geringfügig um 0,5 Prozent überschätzt; dies ist vernachlässigbar. Da in beiden Mikrozensuserhebungen nach dem „gegenwärtigen Schulbesuch" gefragt wird, wäre davon auszugehen, dass Effekte der Fokussierung auf den Interviewzeitpunkt gleichermaßen sowohl 1989 als auch 1991 auftreten. Wenn aber bei den 17-Jährigen

7 Beim Mikrozensus 1996 (einschließlich neue Bundesländer) beträgt die Übererfassung der 17-jährigen Schüler der gymnasialen Oberstufe einschließlich beruflicher Gymnasien rund 23 Prozent.

1989 im Unterschied zu 1991 keine Übererfassung festzustellen ist, kann dies 1989 nicht mit Fokussierungseffekten zusammenhängen. Vergleichbar zum Mikrozensus 1989 enthalten die Zahlen der Bildungsstatistik des Schuljahres 1988/89 neben den 17- bis 20-jährigen Schülern beruflicher Gymnasien die gleichaltrigen Schüler allgemein bildender Schulen, die der gymnasialen Oberstufe zuzurechnen sind. Zwar wird dabei die Sekundarstufe II nur näherungsweise durch die Altersgliederung abgegrenzt, dennoch spricht das Ergebnis dafür, dass die in den Mikrozensen ab 1991 bei den jüngeren, unter 18-jährigen Schülern der gymnasialen Oberstufe festgestellten starken Übererfassungen kaum auf Fokussierungseffekte, sondern auf eine nicht adäquate Unterscheidung der Sekundarstufen I und II zurückzuführen sind.

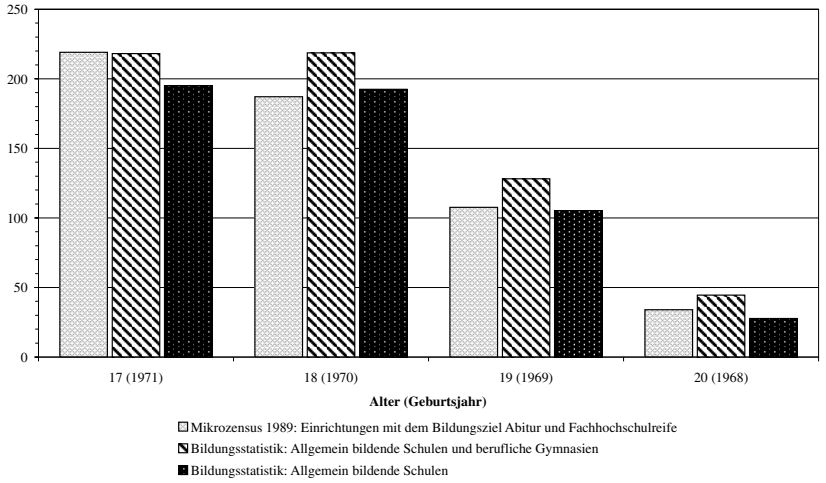


Abb. 3: Schüler allgemein bildender Schulen und beruflicher Gymnasien mit dem Bildungsziel Abitur und Fachhochschulreife im Mikrozensus 1989 und in der Bildungsstatistik (Schuljahr 1988/89) in 1.000

Quellen: Bildungsstatistik: Allgemein bildende Schulen (N=520.400); allgemein bildende Schulen inkl. berufliche Gymnasien (N=609.400); Statistisches Bundesamt 1990.
Mikrozensus 1989 (faktisch anonymisierte 70 %-Substichprobe), Bevölkerung am Ort der Hauptwohnung; hochgerechnete, an die Bevölkerungsfortschreibung angepasste Fallzahlen (N=547.700; ungewichtet n=3.338); eigene Berechnungen.

Die Untererfassung der 17- bis 20-jährigen Schüler allgemein bildender Schulen der gymnasialen Oberstufe und beruflicher Gymnasien im Mikrozensus 1989 um

zehn Prozent ist jedoch nicht zu vernachlässigen. Bei der Frage, womit diese Untererfassung zusammenhängen könnte, ist an die Anmerkungen in Abschnitt 2.3 anzuknüpfen. Wie dort ausgeführt, werden die genauen Zuordnungen der Schulformen zu den Antwortkategorien im Mikrozensus 1989 nur im Interviewerhandbuch genannt. Sowohl im Interviewerbogen als auch im Selbstausfüllerbogen ist lediglich der Sammelbegriff „Gymnasium/Fachoberschule“ angegeben. Nicht explizit genannte Schulen können auf diese Weise leicht übersehen werden. Für die Untererfassung dürfte somit maßgeblich die unvollständige Antwortliste verantwortlich sein.

5 Antwortstabilität der Angaben zum allgemeinen Schulabschluss

Wie wir bisher gesehen haben, treten auch im Mikrozensus systematische Fehler in Form von Klassifikationsfehlern etc. auf. Die Panelangaben des Mikrozensus eröffnen neue Möglichkeiten zur Beurteilung der Antwortkonsistenz bzw. Antwortstabilität im Zeitverlauf. Dabei können geringe Antwortstabilitäten ergänzend zu den vorherigen Analysen auf Probleme bei der Erhebung von Bildungsabschlüssen hinweisen. Trifft die zum Schulbesuch gezogene Schlussfolgerung zu, dass Schüler der gymnasialen Oberstufe und beruflicher Schulen nicht den Definitionen des Mikrozensus entsprechend unterscheidbar sind, ist zu erwarten, dass sich dies auch bei den in diesen Schulformen erworbenen typischen Abschlüssen der allgemeinen bzw. fachgebundenen Hochschulreife (gymnasiale Oberstufe) vs. Fachhochschulreife (berufliche Schulen) widerspiegelt. Aus Mustern von Übergangswahrscheinlichkeiten lässt sich allgemein schließen, dass den Befragten bestimmte Antwortkategorien ähnlich erscheinen. Demnach sind beim Fachhochschulreifeabschluss im Vergleich zu anderen Abschlüssen eine geringere zeitliche Stabilität und eine höhere Übergangswahrscheinlichkeit zum Abitur zu erwarten.

Zur Qualität von Bildungsangaben kann nur auf sehr wenige Erkenntnisse zurückgegriffen werden (Reimer 2001: 114). In der Test-Retest Studie zum ALLBUS 1984 (n=154) wurden drei Interviews im Abstand von vier Wochen durchgeführt. Für den allgemeinen Schulabschluss wurde eine Stabilität von 89 Prozent, für die berufliche Qualifikation nur eine Stabilität von 72 Prozent ermittelt (Porst/Zeifang 1987: 196). Auf der Basis einer zweimaligen Befragung 89 verheirateter und geschiedener Ehepaare im Abstand von etwa einem halben Jahr, mit der sowohl Proxyeffekte als auch Rückerinnerungseffekte untersucht wurden, berichtet Babka von Gostomki (1995) für den Schulabschluss eine Gesamtstabilität von 72 Prozent. Orientiert man sich an diesen zwei Studien, kann eine Antwortstabilität von 70 Prozent als Minimum bezeichnet werden.

Mit dem Mikrozensuspanel liegt für diese Fragestellungen erstmals ein wesentlich umfangreicherer Datensatz vor. Einschränkend muss jedoch bemerkt werden, dass keine Information über die Auskunft gebende Person verfügbar ist. Die Proxy-Angaben wurden zwar erstmals 1999 im Rahmen der EU-Arbeitskräftestichprobe erhoben, jedoch aus Datenschutzgründen nicht in das File aufgenommen. Im Folgenden werden nur räumlich immobile Personen im Alter von 25 bis 50 Jahren ohne Schulbesuch betrachtet, bei denen man von einer abgeschlossenen Schulkarriere ausgehen kann. Mit der oberen Altersgrenze werden der Einfachheit halber die besonderen Aspekte der Freistellung von der Auskunftspflicht für über 51-Jährige ausgeblendet. Ebenfalls vereinfacht die Konzentration auf räumlich Immobile bzw. der Ausschluss wegziehender Personen die Ergebnisdarstellung. Die Selektion dürfte für die Frage konsistenter Antworten unerheblich sein. Für rund 29.000 Personen werden insgesamt rund 87.000 Übergänge beobachtet, d. h. aufgrund von zeitlich inkonsistenten Angaben liegen Mehrfachantworten vor.

Tabelle 1 zeigt für den Abschluss der allgemeinen bzw. fachgebundenen Hochschulreife, dem typischen Abschluss der gymnasialen Oberstufe, eine Antwortstabilität von 85 Prozent. Für Personen mit Fachhochschulreife, die zumeist an beruflichen Schulen erworben wird, wird lediglich eine Wahrscheinlichkeit von 50 Prozent festgestellt, dass dieser Abschluss auch bei der nächsten Befragung genannt wird. Die Übergangswahrscheinlichkeit zur Angabe Abitur beträgt 23 Prozent. Diese überdurchschnittlich inkonsistenten Abschlussangaben korrespondieren mit der oben bereits festgestellten Schwierigkeit von Befragten, den Besuch beruflicher und allgemein bildender Schulen der Sekundarstufe II zu differenzieren.

Hohe Antwortkonsistenzen, d. h. über 80 Prozent liegende Übergangswahrscheinlichkeiten können des Weiteren für die Abschlüsse der Volks- bzw. Hauptschule, der Polytechnischen Oberschule und der Realschule festgestellt werden. Personen, die zunächst angegeben haben, keinen allgemein bildenden Abschluss zu besitzen bzw. ihren Abschluss nicht genannt haben, wechseln im Erhebungszeitraum 1996-1999 häufig zur Angabe eines Hauptschulabschlusses (siehe Kategorien 0, 8 und 9 in Tab. 1). Auch wenn dies wegen den zugrunde liegenden kleinen Fallzahlen nur eingeschränkt zu interpretieren ist, deutet dies darauf hin, dass die Panelangaben zur Korrektur des Item-Nonresponse bzw. zur Überprüfung der Angabe „kein Abschluss“ genutzt werden können.

Tabelle 1: Antwortvariabilitäten zum allgemeinen Schulabschluss im Mikrozensuspanel 1996-1999 (Übergangswahrscheinlichkeiten zwischen den Zeitpunkten *t* und *t+1*) – Zeilenprozentwerte

	Bildungsangabe (*t*)	Bildungsangabe (*t+1*)							
	(*m* Beobachtungen; in Prozent)	0	1	2	3	4	5	8	9
0	Kein Abschluss (1,7 %)	**46,8**	40,0	2,0	2,8	0,3	1,6	1,1	5,5
1	Volks-/Hauptschule (40,3 %)	1,5	**87,4**	1,8	5,3	0,5	0,7	0,5	2,4
2	Polytechnische Oberschule (14,6 %)	0,2	5,9	**83,9**	5,1	1,5	1,5	0,2	1,7
3	Realschule (20,1 %)	0,2	10,0	3,1	**79,0**	3,0	2,4	0,4	1,9
4	Fachhochschulreife (4,2 %)	0,1	4,7	5,3	13,9	**50,5**	23,3	0,3	1,9
5	Allgem./fachgeb. Hochschulreife (15,2 %)	0,2	1,6	1,3	3,2	6,5	**85,2**	0,4	1,7
8	Abschluss vorh., ohne Angabe (0,5 %)	4,4	41,3	5,4	17,6	3,9	9,1	**7,2**	11,1
9	Keine Angabe (3,3 %)	2,8	36,6	8,3	14,4	3,0	9,8	2,6	**22,4**
	Insgesamt (*m* = 87.234; 100 %)	1,6	40,9	14,4	20,4	4,3	15,1	0,5	2,8

Quelle: Faktisch anonymisiertes Mikrozensuspanel 1996-1999; räumlich immobile Personen im Alter von 25 bis 50 Jahren ohne Schulbesuch (n=29.078 Personen).

6 Zusammenfassung

Die am Beispiel des Besuchs der gymnasialen Oberstufe untersuchten Fragen zur Datenqualität haben auf einige Problemstellen aufmerksam gemacht. Das wohl überraschendste Ergebnis besteht in der erheblichen Übererfassung des Besuchs der gymnasialen Oberstufe (Klassenstufe 11-13) durch den Mikrozensus 1996 um insgesamt rund 60 Prozent im Vergleich zur Bildungsstatistik. Diese Übererfassung der Oberstufenschüler ist auf Klassifikationsfehler zurückzuführen, die mit schwierigen, dem Alltagsverständnis nicht direkt entsprechenden Definitionen im Mikrozensus zusammenhängen. Konkret betrifft es die Unterscheidung zwischen der Klassenstufe 11-13 allgemein bildender Schulen („gymnasiale Oberstufe“) einerseits und beruflicher Schulen andererseits, zu denen laut Mikrozensus auch berufliche Gymnasien zählen. Offensichtlich werden die Definitionen und Antwortvorgaben des Mikrozensus von den Befragten nicht vollständig verstanden. Die Gegenüberstellungen mit der Bildungsstatistik deuten darauf hin, dass im Mikrozensus Schüler beruflicher Gymnasien als gymnasiale Oberstufenschüler erfasst sind. Die gravierende Übererfassung der unter 18-jährigen Oberstufenschüler hängt vermutlich damit zusammen, dass die Unterscheidung der Klassenstufen 5-10 vs. 11-13 bzw. der Sekundarstufen I und II nicht klar ist. Zwar

sind diese Befunde teilweise spekulativ, aber auch plausibel. Für alternative Erklärungen und weitere Kontrollen wären zusätzliche Informationen nötig.

Korrespondierend zu der in der Befragungspraxis offenbar schwierigen Unterscheidung zwischen gymnasialer Oberstufe und beruflichen Gymnasien zeigen sich diese Probleme auch in Bezug auf die Konsistenz bzw. Stabilität der Antworten zum allgemeinen Schulabschluss. Für das i. d. R. mit Abschluss der gymnasialen Oberstufe erworbene Abitur kann mit 85 Prozent eine sehr hohe Stabilität festgestellt werden. Dagegen liegt die Stabilität des beruflichen Abschlusses Fachhochschulreife lediglich bei 50 Prozent. Für die anderen Kategorien belegen die ersten hier vorgestellten Ergebnisse des Mikrozensuspanels zur Konsistenz der Antworten zum allgemeinen Schulabschluss erfreulicherweise hohe Antwortstabilitäten von 80 Prozent und mehr.

So ernüchternd die Einbußen der Qualität der Angaben zum Schulbesuch für Bildungsforscher auch sein mögen, ist es gleichwohl für jede sachgerechte Auswertung wichtig zu wissen, in welcher Weise die Daten systematische Fehler aufweisen. Es geht aber nicht nur um die Übereinstimmung mit externen, für valider gehaltenen Datenquellen per se. Für einfache Populationsschätzungen werden die Mikrozensusdaten kaum benötigt, hierfür kann die Bildungsstatistik als Totalerhebung direkt verwendet werden. Die Vorteile des Mikrozensus bestehen vielmehr in der Verfügbarkeit weiterer Merkmale für differenzierte deskriptive Analysen und statistische Modelle. Bei der Verwendung von Bildungsmerkmalen in statistischen Modellen wird infolge fehlender Informationen zur Datenqualität zumeist angenommen, dass die Merkmale frei von systematischen Fehlern sind. In dieser Hinsicht können auf Grundlage der hier berichteten Befunde erste und einfache Konsequenzen gezogen werden. Beispielsweise kann nun bei Analysen zum Besuch der gymnasialen Oberstufe darauf verwiesen werden, dass diese Kategorie nicht nur allgemein bildende Schulen umfasst. Durch die Beschränkung auf über 17-Jährige lässt sich außerdem eine gute Anpassung der Mikrozensusergebnisse an die Bildungsstatistik erreichen, wenn zur gymnasialen Oberstufe auch berufliche Gymnasien gezählt werden (siehe Abbildung 1).

In diesem Zusammenhang der Verwendung bildungsstatistischer Merkmale des Mikrozensus in der Forschung ist darauf hinzuweisen, dass die aktuellen Fragen zum Schulbesuch ab dem Mikrozensus 2005 geändert wurden. Erfragt wird nun der Schulbesuch in den letzten vier Wochen sowie in den letzten zwölf Monaten vor der jeweiligen Befragung. Zwar wird im Fragebogen nach wie vor zwischen dem Besuch allgemein bildender vs. beruflicher Schulen unterschieden, jedoch wird ab 2005 zum Besuch einer beruflichen Schule explizit auch der jeweils erreichbare Abschluss genannt. Ob damit schon die hier festgestellten Fehlklassifikationen der Besucher beruflicher Gymnasien reduziert werden können, bleibt weiteren Analysen vorbehalten. Vor dem Hintergrund, dass der Schulbesuch in den einzelnen Bundesländern unterschiedlich geregelt ist, wird die Erhebung des

Schulbesuchs und von Bildungsabschlüssen nicht nur im Mikrozensus, sondern vermutlich auch in anderen bundesweiten Umfragen auch weiterhin eine Herausforderung darstellen.

Abschließend ist festzuhalten, dass in Deutschland insbesondere im Vergleich zu den in den USA routinemäßig durchgeführten Analysen allgemein zur Datenqualität und besonders zu Messfehlern ein erheblicher Rückstand besteht. Die europäischen statistischen Ämter unternehmen seit einiger Zeit verstärkt Anstrengungen, um diese Defizite aufzuholen (Blanc et al. 2001). Als ein Ergebnis dieser Orientierung auf Aspekte der Datenqualität stellt das Statistische Bundesamt seit kurzem zu verschiedenen Erhebungen und Statistiken Qualitätsberichte zur Verfügung.[8] Aus Sicht der Forschung ist es sehr wünschenswert, dass diesem Schritt weitere, differenziertere Untersuchungen folgen. Die Gründe dafür sind in Empfehlungen an die Datenproduzenten zusammengefasst (Office of Management and Budget 2001: 6-25): „(...) a data user cannot understand the limitations of the data – from a measurement error point of view – unless the data collection program takes steps to explicitly provide such information." Als Nutzer amtlicher Mikrodaten ist aber auch die akademische Forschung gefordert. In dieser Hinsicht wird das Scientific Use File des Mikrozensuspanels – neben neuen substanzwissenschaftlich orientierten Verlaufsanalysen – für Fragen zur Datenqualität eine Vielzahl weiterer methodischer Auswertungsmöglichkeiten bieten.

7 Literaturangaben

Afentakis, A.; Bihler, W. (2005): Das Hochrechnungsverfahren beim unterjährigen Mikrozensus ab 2005. In: Wirtschaft und Statistik, 10, S. 1039-1048.

Babka von Gostomski, C. (1995): Zur Konsistenz und Übereinstimmung von Ehepartnern bei retrospektiv erhobenen Angaben zur Person und zur Beziehung. In: Zeitschrift für Familienforschung 7, 1, S. 6-26.

Bellenberg, G.; Hovestadt, G.; Klemm, K. (2004): Selektivität und Durchlässigkeit im allgemein bildenden Schulsystem. Universität Duisburg-Essen: Arbeitsgruppe Bildungsforschung/Bildungsplanung.

Biemer, P. M.; Lyberg, L. E. (2003): Introduction to Survey Quality. Hoboken, New Jersey: Wiley.

Black, D.; Sanders, S.; Taylor, L. (2003): Measurement of Higher Education in the Census and CPS. In: Journal of the American Statistical Association 98, 463, S. 545-554.

Breiholz, H. (2000): Ergebnisse des Mikrozensus 1999. In: Wirtschaft und Statistik, 5, S. 328-336.

8 Siehe unter www.destatis.de/allg/d/veroe/qualitaetsberichte.htm

Brick, J. M.; Wernimont, J.; Montes, M. (1996): The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component. NCES Working Paper No. 96-14. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Dawe, F.; Knight, I. (1997): A study of proxy response on the Labour Force Survey. In: Survey Methodology Bulletin no. 40: S. 30-36.

Dräther, H.; Fachinger, U., Oelschläger, A. (2001): Selbständige und ihre Altersvorsorge – Möglichkeiten der Analyse anhand der Mikrozensen und erste Ergebnisse. ZeS-Arbeitspapier Nr. 1/01. Universität Bremen: Zentrum für Sozialpolitik.

Esser, H.; Grohmann, H.; Müller, W.; Schäffer, K.-A. (1989): Mikrozensus im Wandel. Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung. In: Statistisches Bundesamt (Hrsg.): Forum der Bundesstatistik, Band 11. Stuttgart: Metzler-Poeschel.

Heidenreich, H.-J. (1994): Hochrechnung des Mikrozensus ab 1990. In: Gabler, S.; Hoffmeyer -Zlotnik, J.; Krebs, D. (Hrsg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 112-123.

KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland] (2001): Grundstruktur des Bildungswesens in der Bundesrepublik Deutschland. Diagramm 1999 für Faltblatt, Sonderdruck und Download. Bonn. URL: www.kmk.org/schul/home.htm - .../doku/ddi-agr.doc.

KMK (2003): Das Bildungswesen in der Bundesrepublik Deutschland 2002. Bonn.
URL: www.kmk.org/dossier/dossier_2002/dossier_2002_dt _ebook.pdf.

Lois, D. (2005): Weiterbildungsbeteiligung älterer Erwerbstätiger – Die Messung im Mikrozensus und der Einfluss soziodemografischer Variablen. Beitrag zur 4. Nutzerkonferenz „Forschung mit dem Mikrozensus: Analysen zur Sozialstruktur und zum Arbeitsmarkt“, Mannheim, 12./13.10.2005.
URL: www.gesis.org/Dauerbeobachtung/GML/Service/Veranstaltungen/4.NK_2005/papers/18_Lois.pdf

Lotze, S.; Breiholz, H. (2002a): Zum neuen Erhebungsdesign des Mikrozensus. Teil 1. In: Wirtschaft und Statistik, 5, S. 359-366.

Lotze, S.; Breiholz, H. (2002b): Zum neuen Erhebungsdesign des Mikrozensus. Teil 2. In: Wirtschaft und Statistik, 6, S. 454-459.

McGovern, P. D.; Bushery, J. M. (1999): Data mining the CPS reinterview: Digging into response error. URL: www.fcsm.gov/99papers/mcgovern.pdf.

Meyer, K. (1994): Zum Auswahlplan des Mikrozensus ab 1990. In: Gabler, S.; Hoffmeyer-Zlotnik, J.; Krebs, D. (Hrsg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 106-111.

NCES [U.S. Department of Education, National Center for Education Statistics] 1997: Measurement Error Studies at the National Center for Education Statistics, NCES 97-464. Washington, DC: NCES.

Office of Management and Budget (2001): Measuring and Reporting Sources of Error in Surveys. Statistical Policy Working Paper 31. URL: www.fcsm.gov/01papers/SPWP31_final.pdf.

Pöschl, H. (1992): Geringfügige Beschäftigung 1990. Ergebnisse des Mikrozensus. In: Wirtschaft und Statistik, 3, S. 166-170.

Porst, R.; Zeifang, K. (1987): A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables. In: Sociological Methods & Research 15, 3, S. 177-218.

Reimer, M. (2001): Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Verlaufsdaten. Kognitive und erhebungspragmatische Aspekte. Materialien aus der Bildungsforschung Nr. 71. Berlin: Max-Planck-Institut für Bildungsforschung.

Riede, T.; Emmerling, D. (1994): Analysen zur Freiwilligkeit der Auskunftserteilung im Mikrozensus. Sind Stichprobenergebnisse bei freiwilliger Auskunftserteilung verzerrt? In: Wirtschaft und Statistik, 9, S. 733-742.

Rendtel, U.; Schimpl-Neimanns, B. (2001): Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. In: ZUMA-Nachrichten 48, S. 85-116.

Rudolph, H. (1998): „Geringfügige Beschäftigung" mit steigender Tendenz. Erhebungskonzepte, Ergebnisse und Interpretationsprobleme der verfügbaren Datenquellen. IAB-Werkstattbericht Nr. 9. Nürnberg: IAB.

Särndal, C.-E.; Swensson, B.; Wretman, J. (1997): Model Assisted Survey Sampling. New York: Springer.

Schimpl-Neimanns, B. (2005): Bildungsverläufe im Mikrozensuspanel 1996-1999: Besuch der gymnasialen Oberstufe bis zum Abitur. ZUMA-Arbeitsbericht 2005/02. Mannheim: ZUMA.

Schupp, J.; Frick, J.; Kaiser, L.; Wagner, G. (1999): Zur Erhebungsproblematik geringfügiger Beschäftigung. Ein Vergleich des Mikrozensus mit dem Sozio-ökonomischen Panel und dem Europäischen Haushaltspanel. In: Lüttinger, P. (Hrsg.): Sozialstrukturanalysen mit dem Mikrozensus. ZUMA-Nachrichten Spezial, Band 6. Mannheim: ZUMA, S. 93-118.

Statistische Ämter des Bundes und der Länder (2004): Ergebnisse des Zensustests. In: Wirtschaft und Statistik, 8, S. 813-833.

Statistisches Bundesamt (1990): Fachserie 11 Bildung und Kultur, Reihe 1 Allgemeinbildende Schulen. Schuljahr 1988/89. Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt (1992): Fachserie 11 Bildung und Kultur, Reihe 1 Allgemeinbildende Schulen. Schuljahr 1990/91. Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt (1996): Fachserie 11 Bildung und Kultur, Reihe 1 Allgemeinbildende Schulen. Schuljahr 1995/96. Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt (1997a): Arbeitsunterlage Allgemeinbildende Schulen Schuljahr 1996/97. Ergänzende Tabellen zur Fachserie 11 Bildung und Kultur, Reihe 1 - Allgemeinbildende Schulen. Wiesbaden.

Statistisches Bundesamt (1997b): Arbeitsunterlage Berufliche Schulen 1995/96. Ergänzende Tabellen zur Fachserie 11 Bildung und Kultur, Reihe 2 - Berufliche Schulen Schuljahr 1995/96.Wiesbaden.

U.S. Census Bureau (2004): Meeting the 21st Century Demographic Data Needs - Implementing the American Community Survey. Report 9: Comparing Social Characteristics With Census 2000. Washington, DC: U.S. Census Bureau.

Weishaupt, H.; Fickermann, D. (2001): Informationelle Infrastruktur im Bereich Bildung und Kultur. Expertise für die Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik. In: Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): Wege zu einer besseren informationellen Infrastruktur. Baden-Baden: Nomos [CD-ROM Beilage zur Buchausgabe].

Verzeichnis der Autorinnen und Autoren

Jun.-Prof. Dr. Nina Baur, Wissenschaftliche Assistentin
Technische Universität Berlin, Fakultät VI: Architektur - Umwelt - Gesellschaft, Institut für Soziologie, FG Methoden soziologischer Forschung, Franklinstraße 28/29, 10587 Berlin

Michael Blohm, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: blohm@zuma-mannheim.de; www.gesis.org/zuma

Prof. Dr. Uwe Engel
Universität Bremen, Institut für emprirische und angewandte Soziologie (EMPAS), Postfach 33 04 40, 28334 Bremen
E-Mail: uengel@empas.uni-bremen.de; www.empas.uni-bremen.de

Prof. Dr. Frank Faulbaum
Universität Duisburg-Essen, Campus Duisburg, LS Sozialwissenschaftliche Methoden/Empirische Sozialforschung, Lotharstr. 65, 47048 Duisburg
E-Mail: faulbaum@uni-duisburg.de; http://soziologie.uni-duisburg.de

PD Dr. Siegfried Gabler, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: gabler@zuma-mannheim.de; www.gesis.org/zuma

Dr. Sabine Häder, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: sabine.haeder@zuma-mannheim.de; www.gesis.org/zuma

PD Dr. Jürgen Hoffmeyer-Zlotnik, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: hoffmeyer-zlotnik@zuma-mannheim.de; www.gesis.org/zuma

Dr. Christian Holst, Director Public Affairs / Politik- & Sozialforschung
Ipsos GmbH, Abtlg. Public Affairs / Politik- & Sozialforschung, Papenkamp 2-6, 23879 Moelln
E-Mail: Christian.Holst@ipsos.de; www.ipsos.de

Ben Jann, Eidgenössische Technische Hochschule Zürich, Soziologie, Scheuchzerstr. 70, 8092 Zürich, Schweiz
E-Mail: ben.jann@soz.gess.ethz.ch

Sonja Krügener
Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen
E-Mail: Sonja.Kruegener@lds.nrw.de

Kersten Magg
Eberhard-Karls-Universität Tübingen, Wirtschaftswissenschaftliche Fakultät, Abteilung Statistik, Ökonometrie und Unternehmensforschung, Mohlstraße 36, Zimmer 322, 72074 Tübingen
E-Mail: kersten.magg@uni-tuebingen.de;
www.uni-tuebingen.de/uni/wwg/ls/assi_magg.htm

Prof. Dr. Ralf Münnich
Universität Trier, Fachbereich IV, Lehrstuhl für Wirtschafts- und Sozialstatistik, Universitätsring 15, 54286 Trier
E-Mail: muennich@uni-trier.de; www.ralf-muennich.de

Bernhard Schimpl-Neimanns, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: schimpl-neimanns@zuma-mannheim.de; www.gesis.org/zuma

Prof. Dr. Rainer Schnell
Universität Konstanz, Fachbereich Politik und Verwaltungswissenschaften, Postfach D92, 78457 Konstanz
E-Mail: rainer.schnell@uni-konstanz.de;
www.uni-konstanz.de/FuF/Verwiss/ Schnell/

Dr. Mark Trappmann
Universität Konstanz, Fachbereich Politik und Verwaltungswissenschaften, Postfach D92, 78457 Konstanz
E-Mail: mark.trappmann@uni-konstanz.de;
www.uni-konstanz.de/struktur/ fuf/polfak/trappmann

PD Dr. Christof Wolf,
Zentrum für Umfragen, Methoden und Analysen (ZUMA), Quadrat B2,1, 68159 Mannheim
E-Mail: wolf@zuma-mannheim.de; www.gesis.org/zuma

Aufgrund der Bedeutung, die Umfragen für die wissenschaftliche Forschung und als Grundlage politischer Entscheidungen haben, ist die Qualität von Umfragedaten ein zentrales Thema der empirischen Sozialforschung. Die in diesem Band versammelten Artikel konzentrieren sich vor allem auf die Qualität der Stichprobe. Im Mittelpunkt der Beiträge stehen unterschiedliche Varianten der Stichprobenziehung, die Entwicklung und Optimierung von Schätzverfahren für verschiedene Erhebungsarten und Erhebungstechnologien sowie die Determinanten systematischer Ausfälle und Möglichkeiten ihrer Reduktion. Die Beiträge sind sowohl für Umfrageforscher als auch für Praktiker, die die Ergebnisse von Umfragen beurteilen müssen, bedeutsam.

Due to the relevance of scientific surveys for science and political decision making alike, the quality of data generated by surveys is a central issue in social research. The contributions in this volume address this issue from various perspectives concentrating on the quality of samples. Sampling methods, optimizing estimation methods for different data collection methods, as well as the determinants of unit non-response and the possibilities of reducing non-response rates are dealt with. The articles in this volume are equally relevant for survey methodologists as for practitioners who have to evaluate the quality of surveys.

InformationsZentrum
Sozialwissenschaften

der Arbeitsgemeinschaft
Sozialwissenschaftlicher Institute e.V.

Lennéstraße 30 • D-53113 Bonn
Telefon 02 28 / 22 81 - 0
Telefax 02 28 / 22 81 - 120

Das IZ ist Mitglied der
Gesellschaft Sozialwissenschaftlicher
Infrastruktureinrichtungen e.V.
Die GESIS ist Mitglied der
Leibniz-Gemeinschaft.

ISBN-10 3-8206-0156-2
ISBN-13 978-3-8206-0156-5
EUR 10,-